

# Mathématiques pour les Médias Numériques

Recueil d'exercices corrigés et aide-mémoire.

Gloria Faccanoni

 <http://faccanoni.univ-tln.fr/enseignements.html>

Année 2017 – 2018

Dernière mise-à-jour : Mercredi 18 octobre 2017

## Table des matières

1	Harmonisation	3
1.1	Espaces vectoriels . . . . .	3
1.2	Éléments d'analyse matricielle . . . . .	5
1.3	Fonctions trigonométriques . . . . .	16
1.4	Transformations élémentaires . . . . .	18
1.5	Nombres complexes . . . . .	20
1.6	Octave/Matlab : guide de survie pour les TP . . . . .	23
	Exercices . . . . .	39
2	Interpolation	83
2.1	Interpolation polynomiale : base canonique, base de LAGRANGE, base de NEWTON . . . . .	83
2.2	Splines : interpolation composite . . . . .	87
2.3	Interpolation Trigonométrique . . . . .	88
	Exercices . . . . .	93
3	Approximation au sens des moindres carrées	111
3.1	Optimisation de fonctions de plusieurs variables . . . . .	111
3.2	Application : fonction de meilleur approximation ( <i>fitting</i> ) . . . . .	120
	Exercices . . . . .	127
4	De l'interpolation à l'approximation d'EDO	157
4.1	Méthodes de quadrature interpolatoires . . . . .	157
4.2	EDO : calcul analytique vs approximation numérique . . . . .	159
	Exercices . . . . .	171
5	Systèmes linéaires	199
5.1	Rappels d'algèbre linéaire . . . . .	199
5.2	Méthodes de résolution analytiques . . . . .	200
5.3	Méthodes de résolution numériques . . . . .	208
	Exercices . . . . .	217
6	Valeurs propres et vecteurs propres	261
6.1	Introduction . . . . .	261
6.2	Localisation des valeurs propres . . . . .	264
6.3	Approximation . . . . .	269
6.4	Décomposition en valeurs singulières . . . . .	269
	Exercices . . . . .	271
	Annales	279

Ce fascicule est un support pour le cours de *mathématiques* de la première année du Diplôme d'ingénieur Cnam – Spécialité INFORMATIQUE – Parcours SCIENCES ET TECHNOLOGIES DES MÉDIAS NUMÉRIQUES (**parcours en alternance**). Ce document donne une aperçue des thèmes qui constituent le socle des connaissances mathématiques indispensables pour votre parcours. On y présente les concepts fondamentaux de la façon la plus intuitive possible avant de procéder à une mise en forme abstraite. Avec un souci de rigueur, mais sans insister sur les concepts les plus abstraits que ne rencontrera probablement pas un élève-ingénieur, on a choisi de développer les preuves les plus utiles. Le moindre calcul est détaillé et les difficultés apparaissent progressivement. Les pré-requis sont limités à ceux acquis en premier cycle. Les exercices et problèmes corrigés, classiques ou plus originaux, sont nombreux et variés. Des exercices nombreux et souvent distrayants éclairent des démonstrations qui vont directement à l'essentiel.

Le but du cours est une ouverture vers des techniques mathématiques appliquées à des problèmes issus des Technologies du numérique. Actuellement il est impossible d'aborder ce sujet sans faire des simulations numériques et le langage Octave/Matlab a été choisi comme langage de programmation du cours. La documentation et les sources peuvent être téléchargées à l'adresse <https://www.gnu.org/software/octave/>. Les notions supposées connues correspondent au programme des cours de Mathématiques (Analyse mathématique des fonctions réelles d'une ou plusieurs variables réelles et Algèbre Linéaire) et Informatiques (Initiation à l'algorithmique).

L'objet de cet aide-mémoire est de proposer une explication succincte des concepts vus en cours. De nombreux livres, parfois très fournis, existent. Ici on a cherché, compte tenu des contraintes de volume horaire, des acquis des étudiants à la première année et des exigences pour la suite du cursus, à dégager les points clés permettant de structurer le travail personnel de l'étudiant voire de faciliter la lecture d'autres ouvrages. Ce polycopié ne dispense pas des séances de cours-TD ni de prendre des notes complémentaires. Il est d'ailleurs important de comprendre et apprendre le cours au fur et à mesure. Ce polycopié est là pour éviter un travail de copie qui empêche parfois de se concentrer sur les explications données oralement mais **ce n'est pas un livre auto-suffisant** et il est loin d'être exhaustif! De plus, ne vous étonnez pas si vous découvrez des erreurs (merci de me les communiquer).

Mathématiques pour l'Informatique		
CM-TD-TP	59h30	17 séances de 3h30

Gloria FACCANONI

IMATH Bâtiment M-117  
Université de Toulon  
Avenue de l'université  
83957 LA GARDE - FRANCE

✉ 0033 (0)4 83 16 66 72

✉ [gloria.faccanoni@univ-tln.fr](mailto:gloria.faccanoni@univ-tln.fr)  
✉ <http://faccanoni.univ-tln.fr>

# Chapitre 1

## Harmonisation

### 1.1 Espaces vectoriels

Dans cette section, nous rappelons les notions élémentaires d'algèbre linéaire que nous utiliserons dans le reste du polycopié.

#### Définition 1.1 (Espace vectoriel)

Un ESPACE VECTORIEL sur un corps  $\mathbb{K}$  ( $\mathbb{K} = \mathbb{C}$  ou  $\mathbb{K} = \mathbb{R}$ ) est un ensemble  $E$  contenant au moins un élément, noté  $\mathbf{0}_E$ , ou simplement  $\mathbf{0}$ , muni d'une loi interne notée  $+$ , appelée *addition*, et d'une loi externe notée  $\cdot$ , appelée *multiplication par un scalaire*, qui possède les propriétés suivantes : pour tout  $\mathbf{u}, \mathbf{v}, \mathbf{w} \in E$  et pour tout  $\alpha, \beta \in \mathbb{K}$ ,

- |   |   |
|---|---|
| ① $\mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{u} + \mathbf{v}) + \mathbf{w}$   | (associativité)                                 |
| ② $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$   | (commutativité)                                 |
| ③ $\mathbf{u} + \mathbf{0}_E = \mathbf{0}_E + \mathbf{u} = \mathbf{u}$  | (existence d'un élément neutre pour l'addition) |
| ④ $\mathbf{u} + (-\mathbf{u}) = (-\mathbf{u}) + \mathbf{u} = \mathbf{0}_E$ en notant $-\mathbf{u} = (-1_{\mathbb{K}}) \cdot \mathbf{u}$ | (existence d'un élément opposé)                 |
| ⑤ $(\alpha + \beta) \cdot \mathbf{u} = \alpha \cdot \mathbf{u} + \beta \cdot \mathbf{u}$  | (compatibilité avec la somme des scalaires)     |
| ⑥ $\alpha \cdot (\mathbf{u} + \mathbf{v}) = \alpha \cdot \mathbf{u} + \alpha \cdot \mathbf{v}$  | (compatibilité avec la somme des vecteurs)      |
| ⑦ $\alpha \cdot (\beta \cdot \mathbf{u}) = (\alpha\beta) \cdot \mathbf{u}$  | (compatibilité avec le produit des scalaires)   |
| ⑧ $1_{\mathbb{K}} \cdot \mathbf{u} = \mathbf{u}$  | (compatibilité avec l'unité)                    |

Les éléments de  $\mathbb{K}$  sont appelés SCALAIRES, ceux de  $E$  sont appelés VECTEURS. L'élément unité de  $\mathbb{K}$  est noté  $1_{\mathbb{K}}$ , l'élément neutre de l'addition  $\mathbf{0}_E$  est appelé VECTEUR NUL, le symétrique d'un vecteur  $\mathbf{u}$  pour l'addition est appelé VECTEUR OPPOSÉ DE  $\mathbf{u}$  et est noté  $-\mathbf{u}$ .

#### EXEMPLE

L'ensemble  $\mathbb{R}^n = \{(x_1, x_2, \dots, x_n) \mid x_i \in \mathbb{R}\}$ ,  $n \geq 1$ , est un espace vectoriel pour les opérations somme  $(x_1, x_2, \dots, x_n) + (y_1, y_2, \dots, y_n) = (x_1 + y_1, x_2 + y_2, \dots, x_n + y_n)$  et multiplication  $\alpha \cdot (x_1, x_2, \dots, x_n) = (\alpha x_1, \alpha x_2, \dots, \alpha x_n)$ .

L'ensemble  $\mathbb{R}_n[x] = \{p(x) = \sum_{i=1}^{n+1} \alpha_i x^{i-1} \mid \alpha_i \in \mathbb{R} \text{ ou } \mathbb{C}\}$  des polynômes de degré inférieur ou égal à  $n$ ,  $n \geq 0$ , à coefficients réels ou complexes, est un espace vectoriel pour les opérations somme  $p_n(x) + q_n(x) = \sum_{i=1}^{n+1} \alpha_i x^{i-1} + \sum_{i=1}^{n+1} \beta_i x^{i-1} = \sum_{i=1}^{n+1} (\alpha_i + \beta_i) x^{i-1}$  et multiplication  $\lambda p_n = \sum_{i=1}^{n+1} (\lambda \alpha_i) x^{i-1}$ .

#### Définition 1.2 (Sous-espace vectoriel)

Soit  $E$  un espace vectoriel. On dit que  $F$  est un SOUS-ESPACE VECTORIEL de  $E$  si et seulement si  $F$  est un espace vectoriel et  $F \subset E$ .

#### EXEMPLE

- ★ L'ensemble  $\{\mathbf{0}_E\}$  constitué de l'unique élément nul est un sous-espace vectoriel de  $E$ , à ne pas confondre avec l'ensemble vide  $\emptyset$  qui n'est pas un sous-espace vectoriel de  $E$  (il ne contient pas le vecteur nul).
- ★ L'ensemble  $E$  est un sous-espace vectoriel de  $E$ .

#### Définition 1.3 (Combinaison linéaire)

Soient  $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_p$  des éléments de l'espace vectoriel  $E$  et  $\alpha_1, \alpha_2, \dots, \alpha_p$  des éléments de  $\mathbb{K}$ . Le vecteur

$$\sum_{i=1}^p \alpha_i \cdot \mathbf{u}_i$$

est appelé COMBINAISON LINÉAIRE des vecteurs  $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_p$ .

### Définition 1.4 (Espace engendré)

Soient  $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_p$  des éléments de l'espace vectoriel  $E$ . L'ensemble de toutes les combinaisons linéaires de ces  $p$  vecteurs fixés est un sous-espace vectoriel de  $E$  appelé SOUS-ESPACE VECTORIEL ENGENDRÉ par  $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_p$  et noté  $\text{Vect}\{\mathbf{u}_1, \dots, \mathbf{u}_p\}$  :

$$\text{Vect}\{\mathbf{u}_1, \dots, \mathbf{u}_p\} = \left\{ \mathbf{u} \in E \mid \exists \alpha_1, \dots, \alpha_p \in \mathbb{R}, \mathbf{u} = \sum_{i=1}^p \alpha_i \cdot \mathbf{u}_i \right\}.$$

Notons que le vecteur  $\mathbf{0}_E$  et les vecteurs  $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_p$  appartiennent à  $\text{Vect}\{\mathbf{u}_1, \dots, \mathbf{u}_p\}$  car pour tout  $j = 1, 2, \dots, p$

$$\mathbf{0}_E = \sum_{i=1}^p 0 \cdot \mathbf{u}_i \quad \text{et} \quad \mathbf{u}_j = \sum_{\substack{i=1 \\ i \neq j}}^p 0 \cdot \mathbf{u}_i + 1 \cdot u_j.$$

### Définition 1.5 (Famille libre, famille génératrice, base)

Soit  $p \in \mathbb{N}^*$ ,  $E$  un espace vectoriel et  $\mathcal{F} = \{\mathbf{u}_1, \dots, \mathbf{u}_p\}$  une famille de vecteurs de  $E$ . On dit que la famille  $\mathcal{F}$  est...

... GÉNÉRATRICE DE  $E$  si et seulement si tout vecteur de  $E$  est combinaison linéaire des éléments de  $F$  :

$$\text{pour tout } \mathbf{u} \in E \text{ il existe } \alpha_1, \dots, \alpha_p \in \mathbb{R} \text{ tel que } \mathbf{u} = \sum_{i=1}^p \alpha_i \cdot \mathbf{u}_i;$$

... LIBRE si et seulement si les  $p$  vecteurs  $\mathbf{u}_1, \dots, \mathbf{u}_p$  sont linéairement indépendants, c'est-dire si

$$\sum_{i=1}^p \alpha_i \cdot \mathbf{u}_i = \mathbf{0}_E \quad \Rightarrow \quad \alpha_i = 0 \ \forall i.$$

Dans le cas contraire la famille est dite liée.

... BASE DE  $E$  si elle est libre et génératrice de  $E$ . Dans ce cas, les réels  $\alpha_1, \dots, \alpha_p$  sont appelées COORDONNÉES ou COMPOSANTES du vecteur  $\mathbf{u}$  dans la base  $\mathcal{F}$ , on écrit  $\text{coord}(\mathbf{u}, \mathcal{F}) = (\alpha_1, \dots, \alpha_p)$  et on dit que  $E$  est de DIMENSION  $p$ . Dans un espace vectoriel  $E$  de dimension finie, toutes les bases ont le même nombre d'éléments. Ce nombre, noté  $\dim(E)$ , est appelé la DIMENSION de  $E$ .

### EXEMPLE

La famille  $\{\mathbf{u} = (1, 0), \mathbf{v} = (0, 1), \mathbf{w} = \mathbf{u} + \mathbf{v}\}$  de vecteurs de  $\mathbb{R}^2$  n'est pas libre : par exemple le vecteur  $(2, -1)$  peut s'écrire comme  $2\mathbf{u} - \mathbf{v}$ , comme  $2\mathbf{w} - 3\mathbf{v}$  etc.

### Théorème 1.6 (de la dimension)

Soit  $\mathcal{F}$  une famille d'éléments de  $E$  de dimension finie  $n$ . Les propriétés suivantes sont équivalentes :

- ①  $\mathcal{F}$  est une base de  $E$
- ②  $\mathcal{F}$  est libre et contient  $n$  éléments
- ③  $\mathcal{F}$  est génératrice de  $E$  et de contient  $n$  éléments
- ④  $\mathcal{F}$  est libre et génératrice de  $E$

### EXEMPLE (BASE CANONIQUE DE $\mathbb{R}^n$ )

Avec  $n \in \mathbb{N}$ , l'espace vectoriel  $\mathbb{R}^n$  est de dimension  $n$ . La famille  $\mathcal{B} = \{(1, 0, \dots, 0); (0, 1, \dots, 0); \dots; (0, 0, \dots, 1)\}$  est une base, appelée BASE CANONIQUE de  $\mathbb{R}^n$ , car pour tout vecteur  $\mathbf{u} \in \mathbb{R}^n$ ,  $\mathbf{u} = (u_1, u_2, \dots, u_n) = u_1 \cdot (1, 0, \dots, 0) + u_2 \cdot (0, 1, \dots, 0) + \dots + u_n \cdot (0, 0, \dots, 1)$  de façon unique.

### EXEMPLE (BASE CANONIQUE DE $\mathbb{R}_n[x]$ )

Avec  $n \in \mathbb{N}$ , l'espace vectoriel  $\mathbb{R}_n[x]$  des polynômes de degré  $\leq n$  est de dimension  $n+1$ . La base  $\mathcal{C} = \{1, x, x^2, \dots, x^n\}$  est appelée BASE CANONIQUE de  $\mathbb{R}_n[x]$  car, pour tout polynôme  $p \in \mathbb{R}_n[x]$ ,  $p(x) = a_0 + a_1 x + a_2 x^2 + \dots + a_n x^n$  de façon unique.

### ATTENTION

Ne pas confondre le vecteur  $\mathbf{u} \in E$  (qui peut être un polynôme, une fonction, une matrice...) avec la matrice colonne de ses coordonnées dans la base  $\mathcal{B}$  de  $E$  (qu'on peut noter  $\text{coord}(\mathbf{u}, \mathcal{B})$ ).

EXEMPLE

Le polynôme  $p(x) = a + bx + cx^2$  a pour coordonnées  $(a, b, c)$  dans la base canonique  $\mathcal{C} = \{1, x, x^2\}$  de  $\mathbb{R}_2[x]$  mais n'est pas égale au vecteur  $(a, b, c)$  de  $\mathbb{R}^3$ . Tous ce qu'on peut dire est que le polynôme  $p(x) = a + bx + cx^2$  de  $\mathbb{R}_2[x]$  et le vecteur  $(a, b, c)$  de  $\mathbb{R}^3$  ont les mêmes coordonnées dans les bases canoniques respectives.

## 1.2 Éléments d'analyse matricielle

On appelle MATRICE  $m \times n$  (ou d'ordre  $m \times n$ ) à coefficients dans  $\mathbb{K}$  tout tableau de  $m$  lignes et  $n$  colonnes d'éléments de  $\mathbb{K}$ . L'ensemble des matrices  $m \times n$  à coefficients dans  $\mathbb{K}$  est noté  $\mathcal{M}_{m,n}(\mathbb{K})$ .

On convient de noter  $a_{ij}$  l'élément de la matrice situé sur la  $i$ -ème ligne et  $j$ -ème colonne ( $1 \leq i \leq m$  et  $1 \leq j \leq n$ ).

Une matrice  $\mathbb{A}$  est représentée entre deux parenthèses ou deux crochets :

$$\mathbb{A} = \begin{pmatrix} a_{11} & \dots & a_{1j} & \dots & a_{1n} \\ \vdots & & \vdots & & \vdots \\ a_{i1} & \dots & a_{ij} & \dots & a_{in} \\ \vdots & & \vdots & & \vdots \\ a_{m1} & \dots & a_{mj} & \dots & a_{mn} \end{pmatrix} \quad \text{ou} \quad \mathbb{A} = \begin{bmatrix} a_{11} & \dots & a_{1j} & \dots & a_{1n} \\ \vdots & & \vdots & & \vdots \\ a_{i1} & \dots & a_{ij} & \dots & a_{in} \\ \vdots & & \vdots & & \vdots \\ a_{m1} & \dots & a_{mj} & \dots & a_{mn} \end{bmatrix}$$

ou encore

$$\mathbb{A} = (a_{ij})_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}} \quad \text{ou} \quad \mathbb{A} = [a_{ij}]_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}}$$

EXEMPLE

La matrice  $\mathbb{A} = \begin{pmatrix} -1 & 4 & 2 \\ 0 & 1 & -3 \\ 4 & 1 & 5 \end{pmatrix}$  est carrée et d'ordre 3 à coefficients dans  $\mathbb{Z}$ .

- ★ Si  $m = n$  on dit qu'on a une MATRICE CARRÉE. L'ensemble des matrices carrées d'ordre  $n$  à coefficients dans  $\mathbb{K}$  est noté  $\mathcal{M}_n(\mathbb{K})$ .
- ★ Une matrice  $m \times 1$  est appelée VECTEUR-COLONNE et une matrice  $1 \times n$  est appelée VECTEUR-LIGNE.
- ★ La MATRICE NULLE, notée  $\mathbb{O}_{m,n}$ , est la matrice dont tous les éléments sont nuls.
- ★ On appelle MATRICE DIAGONALE toute matrice carrée  $\mathbb{D} = (d_{ij})_{1 \leq i,j \leq n}$  telle que  $i \neq j \implies d_{ij} = 0$ . Si on note  $d_i \equiv d_{ii}$ , une matrice diagonale est de la forme

$$\mathbb{D}_n = \begin{pmatrix} d_1 & 0 & \dots & 0 & 0 \\ 0 & d_2 & \dots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & d_{n-1} & 0 \\ 0 & 0 & \dots & 0 & d_n \end{pmatrix}.$$

On la note  $\text{Diag}(d_1, d_2, \dots, d_n)$ .

- ★ La MATRICE IDENTITÉ d'ordre  $n$ , notée  $\mathbb{I}_n$ , est la matrice diagonale  $\text{Diag}(1, 1, \dots, 1)$ .

- ★ On dit qu'une matrice carrée  $\mathbb{A} = (a_{ij})_{1 \leq i,j \leq n}$  est

- ★ TRIANGULAIRE SUPÉRIEURE si  $i > j \implies a_{ij} = 0$ ,
- ★ TRIANGULAIRE INFÉRIEURE si  $i < j \implies a_{ij} = 0$ .

Une matrice triangulaire supérieure et inférieure (i.e.  $i \neq j \implies a_{ij} = 0$ ) est une matrice diagonale.

EXEMPLE

$\mathbb{U}$  est une matrice triangulaire supérieure,  $\mathbb{L}$  une matrice triangulaire inférieure,  $\mathbb{D}$  une matrice diagonale et  $\mathbb{I}_4$  la matrice identité d'ordre 4 :

$$\mathbb{U} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 5 & 6 & 7 \\ 0 & 0 & 2 & -1 \\ 0 & 0 & 0 & -5 \end{pmatrix} \quad \mathbb{L} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 \\ 5 & -1 & 2 & 0 \\ 7 & 9 & 15 & 4 \end{pmatrix} \quad \mathbb{D} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -8 & 0 & 0 \\ 0 & 0 & 7 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \mathbb{I}_4 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

 **Définition 1.7 (Addition de matrices)**

Si  $\mathbb{A} = (a_{ij})_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}}$  et  $\mathbb{B} = (b_{ij})_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}}$  sont deux matrices  $m \times n$ , on définit l'ADDITION des matrices par

$$\mathbb{A} + \mathbb{B} = (a_{ij} + b_{ij})_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}}.$$

La MATRICE OPPOSÉE D'UNE MATRICE  $\mathbb{A}$  est notée  $-\mathbb{A}$ . Si  $\mathbb{A} = (a_{ij})_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}}$  alors  $-\mathbb{A} = (-a_{ij})_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}}$ .

 EXEMPLE

Soient les matrices  $2 \times 3$  suivantes :

$$\mathbb{A} = \begin{pmatrix} 3 & 4 & 2 \\ 1 & 3 & 5 \end{pmatrix}, \quad \mathbb{B} = \begin{pmatrix} 6 & 1 & 9 \\ 2 & 0 & 3 \end{pmatrix}.$$

La somme de  $\mathbb{A}$  et  $\mathbb{B}$  est la matrice  $2 \times 3$  suivante :

$$\mathbb{A} + \mathbb{B} = \begin{pmatrix} 3+6 & 4+1 & 2+9 \\ 1+2 & 3+0 & 5+3 \end{pmatrix} = \begin{pmatrix} 9 & 5 & 11 \\ 3 & 3 & 8 \end{pmatrix}.$$

 ATTENTION

La somme de deux matrices d'ordres différents n'est pas définie.

 **Propriété 1.8**

Si  $\mathbb{A}$ ,  $\mathbb{B}$  et  $\mathbb{C}$  sont des matrices de même ordre, alors nous avons

- ★  $\mathbb{A} + \mathbb{B} = \mathbb{B} + \mathbb{A}$  (commutativité),
- ★  $\mathbb{A} + (\mathbb{B} + \mathbb{C}) = (\mathbb{A} + \mathbb{B}) + \mathbb{C}$  (associativité).

 EXEMPLE

Soient les matrices  $2 \times 2$  suivantes :

$$\mathbb{A} = \begin{pmatrix} 1 & -1 \\ 3 & 0 \end{pmatrix}, \quad \mathbb{B} = \begin{pmatrix} 6 & -5 \\ 2 & 1 \end{pmatrix}, \quad \mathbb{C} = \begin{pmatrix} 0 & 2 \\ 2 & 4 \end{pmatrix}.$$

On a alors

$$\mathbb{A} + \mathbb{B} = \begin{pmatrix} 1+6 & -1-5 \\ 3+2 & 0+1 \end{pmatrix} = \begin{pmatrix} 7 & -6 \\ 5 & 1 \end{pmatrix}, \quad \mathbb{B} + \mathbb{A} = \begin{pmatrix} 6+1 & -5-1 \\ 2+3 & 1+0 \end{pmatrix} = \begin{pmatrix} 7 & -6 \\ 5 & 1 \end{pmatrix}, \quad \mathbb{B} + \mathbb{C} = \begin{pmatrix} 6+0 & -5+2 \\ 2+2 & 1+4 \end{pmatrix} = \begin{pmatrix} 6 & -3 \\ 4 & 5 \end{pmatrix}.$$

De plus,

$$(\mathbb{A} + \mathbb{B}) + \mathbb{C} = \begin{pmatrix} 7 & -4 \\ 7 & 5 \end{pmatrix}, \quad \mathbb{A} + (\mathbb{B} + \mathbb{C}) = \begin{pmatrix} 7 & -4 \\ 7 & 5 \end{pmatrix}.$$

 **Définition 1.9 (Produit d'une matrice par un scalaire)**

Si  $\mathbb{A} = (a_{ij})_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}}$  est une matrice  $m \times n$  et si  $\alpha \in \mathbb{K}$ , on définit le PRODUIT D'UNE MATRICE PAR UN SCALAIRE par

$$\alpha \cdot \mathbb{A} = (\alpha \cdot a_{ij})_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}}$$

 **Propriété 1.10**

Soient  $\mathbb{A}$  et  $\mathbb{B}$  deux matrices de même ordre et  $\alpha \in \mathbb{K}$  un scalaire, on a

- ★  $\alpha \cdot (\mathbb{A} + \mathbb{B}) = \alpha \cdot \mathbb{A} + \alpha \cdot \mathbb{B}$  (distributivité).

 EXEMPLE

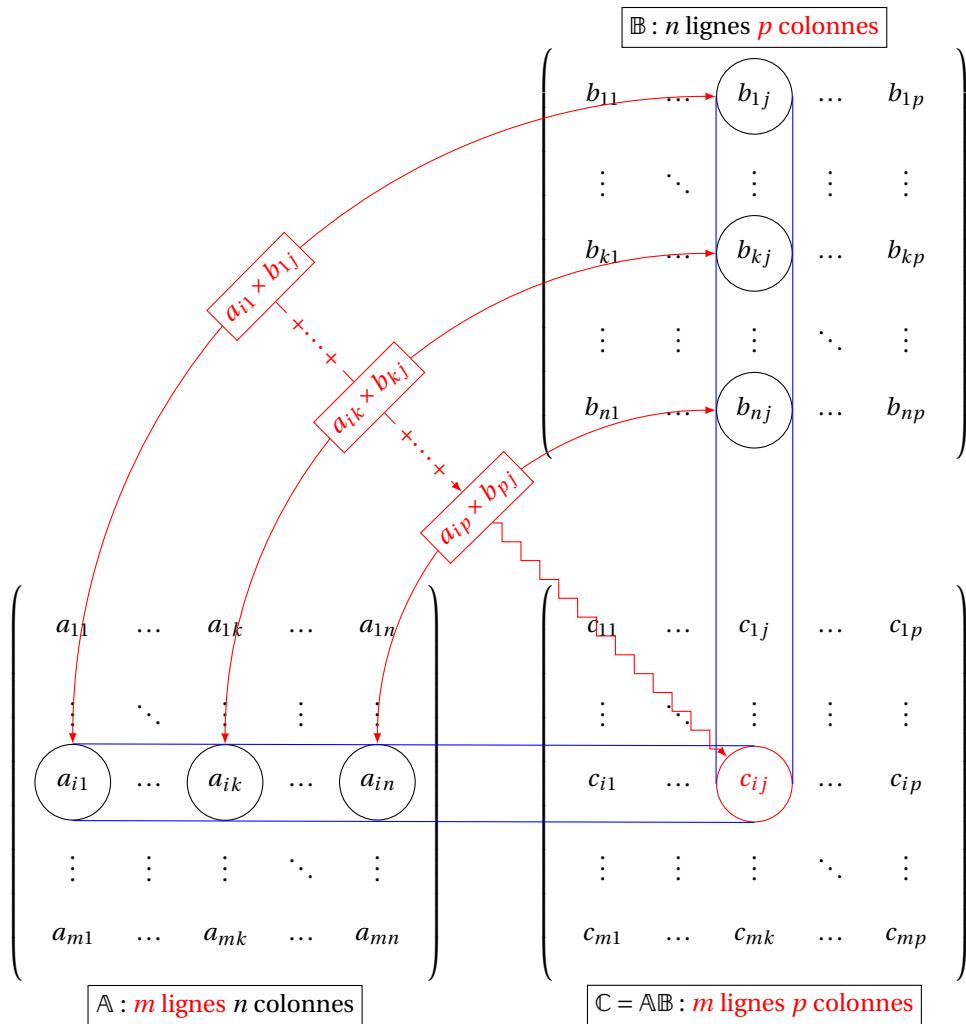
Soit  $\alpha = \frac{1}{2}$  et  $\mathbb{A} = \begin{pmatrix} 3 & 4 & 2 \\ 1 & 3 & 5 \end{pmatrix}$ . Alors  $\alpha \cdot \mathbb{A} = \begin{pmatrix} \frac{3}{2} & 2 & 1 \\ \frac{1}{2} & \frac{3}{2} & \frac{5}{2} \end{pmatrix}$ .

### Définition 1.11 (Produit de matrices)

Si  $\mathbb{A} = (a_{ik})_{\substack{1 \leq i \leq m \\ 1 \leq k \leq n}}$  est une matrice  $m \times n$  et  $\mathbb{B} = (b_{kj})_{\substack{1 \leq k \leq n \\ 1 \leq j \leq p}}$  une matrice  $n \times p$ , on définit le PRODUIT DES MATRICES par

$$\mathbb{AB} = \left( \sum_{k=1}^n a_{ik} b_{kj} \right)_{\substack{1 \leq i \leq m \\ 1 \leq j \leq p}}$$

C'est une matrice  $m \times p$ .



### EXEMPLE

Soient les deux matrices

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

La matrice  $\mathbb{A}$  est d'ordre  $2 \times 3$ , la matrice  $\mathbb{B}$  est d'ordre  $3 \times 3$ , donc la matrice produit  $\mathbb{AB}$  est une matrice d'ordre  $2 \times 3$  :

$$\mathbb{AB} = \begin{pmatrix} 1 \times 1 + 3 \times 0 + 0 \times 0 & 1 \times 2 + 3 \times 2 + 0 \times (-1) & 1 \times 0 + 3 \times 3 + 0 \times (-2) \\ -1 \times 1 + 1 \times 0 + 2 \times 0 & -1 \times 2 + 1 \times 2 + 2 \times (-1) & -1 \times 0 + 1 \times 3 + 2 \times (-2) \end{pmatrix} = \begin{pmatrix} 1 & 7 & 9 \\ -1 & -2 & -1 \end{pmatrix}.$$

### Propriété 1.12

Si  $\mathbb{A} \in \mathcal{M}_{m,n}(\mathbb{K})$ ,  $\mathbb{B} \in \mathcal{M}_{n,p}(\mathbb{K})$ ,  $\mathbb{C} \in \mathcal{M}_{p,q}(\mathbb{K})$ , alors

- \*  $\mathbb{A}(\mathbb{BC}) = (\mathbb{AB})\mathbb{C}$  (associativité) ;
- \*  $\mathbb{A}(\mathbb{B} + \mathbb{C}) = \mathbb{AB} + \mathbb{AC}$  (distributivité) ;

Si  $\mathbb{A} \in \mathcal{M}_n(\mathbb{K})$  alors  $\mathbb{A}\mathbb{I}_n = \mathbb{I}_n\mathbb{A} = \mathbb{A}$ .

 **ATTENTION**

$\mathbb{A}\mathbb{B} \neq \mathbb{B}\mathbb{A}$  en général (non commutativité).

Prenons le cas général avec  $\mathbb{A}$  d'ordre  $m \times p$  et  $\mathbb{B}$  d'ordre  $p \times n$ . Le produit  $\mathbb{A}\mathbb{B}$  est défini, c'est une matrice d'ordre  $m \times n$ . Qu'en est-il du produit  $\mathbb{B}\mathbb{A}$ ? Il faut distinguer trois cas :

- \* si  $m \neq n$  le produit  $\mathbb{B}\mathbb{A}$  n'est pas défini;
- \* si  $m = n$  mais  $p \neq n$ , le produit  $\mathbb{A}\mathbb{B}$  est défini et c'est une matrice d'ordre  $m \times n$  tandis que le produit  $\mathbb{B}\mathbb{A}$  est défini mais c'est une matrice d'ordre  $p \times p$  donc  $\mathbb{A}\mathbb{B} \neq \mathbb{B}\mathbb{A}$ ;
- \* si  $m = n = p$ ,  $\mathbb{A}$  et  $\mathbb{B}$  sont deux matrices carrées d'ordre  $m$ . Les produits  $\mathbb{A}\mathbb{B}$  et  $\mathbb{B}\mathbb{A}$  sont aussi carrés et d'ordre  $m$  mais là encore, en général,  $\mathbb{A}\mathbb{B} \neq \mathbb{B}\mathbb{A}$ ;

 **EXEMPLE**

Soient les matrices

$$\mathbb{A} = \begin{pmatrix} 1 & -1 \\ 3 & 0 \end{pmatrix}, \quad \mathbb{B} = \begin{pmatrix} 6 & -5 \\ 2 & 1 \end{pmatrix}.$$

On obtient

$$\mathbb{A}\mathbb{B} = \begin{pmatrix} 4 & -6 \\ 18 & -15 \end{pmatrix} \quad \text{et} \quad \mathbb{B}\mathbb{A} = \begin{pmatrix} -9 & -6 \\ 5 & -2 \end{pmatrix}.$$

 **Définition 1.13 (Matrice TRANSPOSÉE)**

Si  $\mathbb{A} = (a_{ij})_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}}$  est une matrice  $\mathcal{M}_{m,n}(\mathbb{R})$ , on définit la matrice TRANSPOSÉE de  $\mathbb{A}$ , notée  $\mathbb{A}^T$ , par  $\mathbb{A}^T = (a_{ji})_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}}$ . C'est donc une matrice de  $\mathcal{M}_{n,m}(\mathbb{R})$  obtenue en échangeant lignes et colonnes de la matrice initiale.

 **Définition 1.15 (Matrice ADJOINTE)**

Si  $\mathbb{A} = (a_{ij})_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}}$  est une matrice  $\mathcal{M}_{m,n}(\mathbb{C})$ , on définit la matrice ADJOINTE (ou CONJUGUÉE TRANSPOSÉE) de  $\mathbb{A}$ , notée  $\mathbb{A}^H$ , par  $\mathbb{A}^H = (\bar{a}_{ji})_{\substack{1 \leq j \leq n \\ 1 \leq i \leq m}}$ . C'est donc une matrice de  $\mathcal{M}_{n,m}(\mathbb{C})$  obtenue en échangeant lignes et colonnes de la matrice initiale et en prenant le nombre complexe conjugué.

 **Propriété 1.14**

- \*  $(\mathbb{A}^T)^T = \mathbb{A}$  si  $\mathbb{A} \in \mathcal{M}_{m,n}(\mathbb{K})$ ,
- \*  $(\alpha\mathbb{A})^T = \alpha\mathbb{A}^T$  si  $\alpha \in \mathbb{K}$  et  $\mathbb{A} \in \mathcal{M}_{m,n}(\mathbb{K})$ ,
- \*  $(\mathbb{A} + \mathbb{B})^T = \mathbb{A}^T + \mathbb{B}^T$  si  $\mathbb{A}, \mathbb{B} \in \mathcal{M}_{m,n}(\mathbb{K})$ ,
- \*  $(\mathbb{A}\mathbb{B})^T = \mathbb{B}^T\mathbb{A}^T$  si  $\mathbb{A} \in \mathcal{M}_{m,n}(\mathbb{K})$  et  $\mathbb{B} \in \mathcal{M}_{n,p}(\mathbb{K})$ .

 **Propriété 1.16**

- \*  $(\mathbb{A}^H)^H = \mathbb{A}$  si  $\mathbb{A} \in \mathcal{M}_{m,n}(\mathbb{C})$ ,
- \*  $(\alpha\mathbb{A})^H = \bar{\alpha}\mathbb{A}^H$  si  $\alpha \in \mathbb{C}$  et  $\mathbb{A} \in \mathcal{M}_{m,n}(\mathbb{C})$ ,
- \*  $(\mathbb{A} + \mathbb{B})^H = \mathbb{A}^H + \mathbb{B}^H$  si  $\mathbb{A}, \mathbb{B} \in \mathcal{M}_{m,n}(\mathbb{C})$ ,
- \*  $(\mathbb{A}\mathbb{B})^H = \mathbb{B}^H\mathbb{A}^H$  si  $\mathbb{A} \in \mathcal{M}_{m,n}(\mathbb{C})$  et  $\mathbb{B} \in \mathcal{M}_{n,p}(\mathbb{C})$ .

 **EXEMPLE**

Soit la matrice  $\mathbb{A}$  d'ordre  $2 \times 3$  suivante

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

Sa transposée est la matrice  $\mathbb{A}^T$  d'ordre  $3 \times 2$  suivante

$$\mathbb{A}^T = \begin{pmatrix} 1 & 3 \\ -1 & 0 \\ 5 & 7 \end{pmatrix}.$$

 **Définition 1.17 (Matrice matrice symétrique)**

Une matrice  $\mathbb{A}$  est dite SYMÉTRIQUE si  $\mathbb{A}^T = \mathbb{A}$ , i.e. si  $a_{ij} = a_{ji}$  pour tout  $i \neq j$ .

 **Définition 1.18 (Matrice hermitienne ou autoadjointe)**

Une matrice  $\mathbb{A}$  est dite HERMITIENNE si  $\mathbb{A}^H = \mathbb{A}$ , i.e. si  $\bar{a}_{ij} = a_{ji}$  pour tout  $i \neq j$ .

 **EXEMPLE**

La matrice  $\mathbb{A} = \begin{pmatrix} 1 & 5 & -9 \\ 5 & 4 & 0 \\ -9 & 0 & 7 \end{pmatrix}$  est symétrique.

### Définition 1.19 (Matrice INVERSIBLE, matrice SINGULIÈRE)

Une matrice carrée  $\mathbb{A} \in \mathcal{M}_n(\mathbb{K})$  est dite INVERSIBLE (ou régulière) si elle est symétrisable pour le produit matriciel, autrement dit s'il existe une matrice  $\mathbb{B} \in \mathcal{M}_n(\mathbb{K})$  telle que

$$\mathbb{A}\mathbb{B} = \mathbb{B}\mathbb{A} = \mathbb{I}_n.$$

Si une telle matrice existe, on la note  $\mathbb{A}^{-1}$  et on l'appelle matrice INVERSE de  $\mathbb{A}$ .

Une matrice non inversible est dite SINGULIÈRE.

### Proposition 1.20

Soit  $\mathbb{A}$  et  $\mathbb{B}$  deux matrices inversibles, alors

- \*  $\mathbb{A}^{-1}$  l'est aussi et  $(\mathbb{A}^{-1})^{-1} = \mathbb{A}$ ,
- \*  $\mathbb{A}^T$  l'est aussi et  $(\mathbb{A}^T)^{-1} = (\mathbb{A}^{-1})^T$ ,
- \*  $\mathbb{A}\mathbb{B}$  l'est aussi et  $(\mathbb{A}\mathbb{B})^{-1} = \mathbb{B}^{-1}\mathbb{A}^{-1}$ .

### Définition 1.21 (Matrice ORTHOGONALE)

Une matrice carrée  $\mathbb{A} \in \mathcal{M}_n(\mathbb{K})$  est dite ORTHOGONALE si elle est inversible et  $\mathbb{A}^T\mathbb{A} = \mathbb{A}\mathbb{A}^T = \mathbb{I}_n$ , i.e. si

$$\mathbb{A}^T = \mathbb{A}^{-1}.$$

### Définition 1.22 (Matrice UNITAIRE)

Une matrice carrée  $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$  est dite UNITAIRE si elle est inversible et  $\overline{\mathbb{A}}^H\mathbb{A} = \mathbb{A}\overline{\mathbb{A}}^H = \mathbb{I}_n$ , i.e. si

$$\mathbb{A}^H = \mathbb{A}^{-1}.$$

### Définition 1.23 (TRACE d'une matrice)

Si  $\mathbb{A}$  est une matrice carrée d'ordre  $n$ , on définit la TRACE de  $\mathbb{A}$  comme la somme des éléments de la diagonale principale :

$$\text{tr}(\mathbb{A}) \equiv \sum_{i=1}^n a_{ii}.$$

#### EXEMPLE

La trace de la matrice  $\mathbb{A} = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 2 & 3 \\ 0 & -1 & -2 \end{pmatrix}$  est  $\text{tr}(\mathbb{A}) = a_{11} + a_{22} + a_{33} = 1 + 2 + (-2) = 1$ .

### Propriété 1.24

Si  $\mathbb{A}$  et  $\mathbb{B}$  sont deux matrices carrées d'ordre  $n$ , alors

- \*  $\text{tr}(\mathbb{A}^T) = \text{tr}(\mathbb{A})$ ,
- \*  $\text{tr}(\mathbb{A} + \mathbb{B}) = \text{tr}(\mathbb{A}) + \text{tr}(\mathbb{B})$ .

Si  $\mathbb{A}$  est une matrice  $m \times n$  et  $\mathbb{B}$  une matrice  $n \times m$ , alors

- \*  $\text{tr}(\mathbb{A}\mathbb{B}) = \text{tr}(\mathbb{B}\mathbb{A})$ .

## 1.2.1 Définition et calcul pratique d'un déterminant

### Définition 1.25 (DÉTERMINANT d'une matrice d'ordre $n$ (règle de LAPLACE))

Soit  $\mathbb{A}$  une matrice carrée d'ordre  $n$ .

Étant donné un couple  $(i, j)$  d'entiers,  $1 \leq i, j \leq n$ , on note  $\mathbb{A}_{ij}$  la matrice carrée d'ordre  $n - 1$  obtenue en supprimant la  $i$ -ème ligne et la  $j$ -ème colonne de  $\mathbb{A}$ .

Le DÉTERMINANT de  $\mathbb{A}$ , noté  $\det(\mathbb{A})$  ou  $|\mathbb{A}|$ , est défini par récurrence sur l'ordre de la matrice  $\mathbb{A}$  :

- \* si  $n = 1$  : le déterminant de  $\mathbb{A}$  est le nombre

$$\det(\mathbb{A}) \equiv a_{11},$$

- \* si  $n > 1$  : le déterminant de  $\mathbb{A}$  est le nombre

$$\det(\mathbb{A}) \equiv \sum_{j=1}^n (-1)^{i+j} a_{ij} \det(\mathbb{A}_{ij}) \quad \text{quelque soit la ligne } i, 1 \leq i \leq n,$$

ou, de manière équivalente, le nombre

$$\det(\mathbb{A}) \equiv \sum_{i=1}^n (-1)^{i+j} a_{ij} \det(\mathbb{A}_{ij}) \quad \text{quelque soit la colonne } j, 1 \leq j \leq n.$$

### Astuce

Pour se souvenir des signes de ces deux formules, on peut remarquer que la distribution des signes + et - avec la formule  $(-1)^{i+j}$  est analogue à la distribution des cases noirs et blanches sur un damier :

$$\begin{vmatrix} + & - & + & - & \dots \\ - & + & - & + & \dots \\ + & - & + & - & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{vmatrix}$$

### EXEMPLE

Soit la matrice

$$\mathbb{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

alors

$$\det(\mathbb{A}_{11}) = a_{22}, \quad \det(\mathbb{A}_{12}) = a_{21}, \quad \det(\mathbb{A}_{21}) = a_{12}, \quad \det(\mathbb{A}_{22}) = a_{11},$$

donc on peut calculer  $\det(\mathbb{A})$  par l'une des formules suivantes :

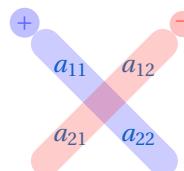
- \*  $a_{11} \det(\mathbb{A}_{11}) - a_{12} \det(\mathbb{A}_{12}) = a_{11}a_{22} - a_{12}a_{21}$  (développement suivant la ligne  $i = 1$ )
- \*  $-a_{21} \det(\mathbb{A}_{21}) + a_{22} \det(\mathbb{A}_{22}) = -a_{21}a_{12} + a_{22}a_{11}$  (développement suivant la ligne  $i = 2$ )
- \*  $a_{11} \det(\mathbb{A}_{11}) - a_{21} \det(\mathbb{A}_{21}) = a_{11}a_{22} - a_{21}a_{12}$  (développement suivant la colonne  $j = 1$ )
- \*  $-a_{12} \det(\mathbb{A}_{12}) + a_{22} \det(\mathbb{A}_{22}) = -a_{12}a_{21} + a_{22}a_{11}$  (développement suivant la colonne  $j = 2$ )

Ces formules donnent bien le même résultat.

### Déterminant d'une matrice d'ordre 2 — méthode pratique

Soit  $\mathbb{A}$  une matrice carrée d'ordre  $n = 2$ .

$$\det(\mathbb{A}) = \det \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = a_{11}a_{22} - a_{12}a_{21}.$$



### EXEMPLE

$$\det \begin{pmatrix} 5 & 7 \\ 4 & 3 \end{pmatrix} = 5 \times 3 - 7 \times 4 = 15 - 28 = -13.$$

### EXEMPLE

Soit la matrice

$$\mathbb{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

alors

$$\begin{aligned} \det(\mathbb{A}_{11}) &= \det \begin{pmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{pmatrix} = a_{22}a_{33} - a_{23}a_{32}, & \det(\mathbb{A}_{12}) &= \det \begin{pmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{pmatrix} = a_{21}a_{33} - a_{23}a_{31}, \\ \det(\mathbb{A}_{13}) &= \det \begin{pmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix} = a_{21}a_{32} - a_{22}a_{31}, & \det(\mathbb{A}_{21}) &= \det \begin{pmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{pmatrix} = a_{12}a_{33} - a_{13}a_{32}, \\ \det(\mathbb{A}_{22}) &= \det \begin{pmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{pmatrix} = a_{11}a_{33} - a_{13}a_{31}, & \det(\mathbb{A}_{23}) &= \det \begin{pmatrix} a_{11} & a_{12} \\ a_{31} & a_{32} \end{pmatrix} = a_{11}a_{32} - a_{12}a_{31}, \end{aligned}$$

$$\det(\mathbb{A}_{31}) = \det \begin{pmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{pmatrix} = a_{12}a_{23} - a_{13}a_{22}, \quad \det(\mathbb{A}_{32}) = \det \begin{pmatrix} a_{11} & a_{13} \\ a_{21} & a_{23} \end{pmatrix} = a_{11}a_{23} - a_{13}a_{21},$$

$$\det(\mathbb{A}_{33}) = \det \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = a_{11}a_{22} - a_{12}a_{21},$$

donc on peut calculer  $\det(\mathbb{A})$  par l'une des formules suivantes :

- ★  $a_{11}\det(\mathbb{A}_{11}) - a_{12}\det(\mathbb{A}_{12}) + a_{13}\det(\mathbb{A}_{13})$  (développement suivant la ligne  $i = 1$ )
- ★  $-a_{21}\det(\mathbb{A}_{21}) + a_{22}\det(\mathbb{A}_{22}) - a_{23}\det(\mathbb{A}_{23})$  (développement suivant la ligne  $i = 2$ )
- ★  $a_{31}\det(\mathbb{A}_{31}) - a_{32}\det(\mathbb{A}_{32}) + a_{33}\det(\mathbb{A}_{33})$  (développement suivant la ligne  $i = 3$ )
- ★  $-a_{11}\det(\mathbb{A}_{11}) + a_{21}\det(\mathbb{A}_{21}) - a_{31}\det(\mathbb{A}_{31})$  (développement suivant la colonne  $j = 1$ )
- ★  $a_{12}\det(\mathbb{A}_{12}) - a_{22}\det(\mathbb{A}_{22}) + a_{32}\det(\mathbb{A}_{32})$  (développement suivant la colonne  $j = 2$ )
- ★  $-a_{13}\det(\mathbb{A}_{13}) + a_{23}\det(\mathbb{A}_{23}) - a_{33}\det(\mathbb{A}_{33})$  (développement suivant la colonne  $j = 3$ )

Quelques calculs montrent que ces formules donnent bien le même résultat.

### Propriété 1.26

1. Le déterminant d'une matrice triangulaire est égal au produit des éléments diagonaux.
2. Le déterminant d'une matrice orthogonale est égal à 1.

### Astuce

Il convient d'utiliser la définition de déterminant après avoir fait apparaître sur une même rangée le plus possible de zéro sachant que

- ★ si deux colonnes (resp. deux lignes) sont identiques ou proportionnelles, alors  $\det(\mathbb{A}) = 0$  ;
- ★ si on multiplie une colonne (resp. une ligne) par un scalaire  $\alpha \neq 0$ , alors le déterminant est multiplié par  $\alpha$  ;
- ★ si on échange deux colonnes (resp. deux lignes), alors le déterminant est changé en son opposé (*i.e.*, le déterminant change de signe) ;
- ★ on ne change pas un déterminant si on ajoute à une colonne (resp. une ligne) une combinaison linéaire des autres colonnes (resp. lignes), *i.e.*

$$C_i \leftarrow C_i + \alpha C_j, \quad L_i \leftarrow L_i + \alpha L_j,$$

avec  $j \neq i$  et  $\alpha \neq 0$ .

### EXEMPLE

Soit la matrice

$$\mathbb{A} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 2 & 0 \\ 0 & 3 & 5 \end{pmatrix}$$

alors

$$\begin{aligned} \det(\mathbb{A}_{11}) &= \det \begin{pmatrix} 2 & 0 \\ 3 & 5 \end{pmatrix} = 10, & \det(\mathbb{A}_{12}) &= \det \begin{pmatrix} 0 & 0 \\ 0 & 5 \end{pmatrix} = 0, & \det(\mathbb{A}_{13}) &= \det \begin{pmatrix} 0 & 2 \\ 0 & 3 \end{pmatrix} = 0, \\ \det(\mathbb{A}_{21}) &= \det \begin{pmatrix} 0 & 1 \\ 3 & 5 \end{pmatrix} = -3, & \det(\mathbb{A}_{22}) &= \det \begin{pmatrix} 1 & 1 \\ 0 & 5 \end{pmatrix} = 5, & \det(\mathbb{A}_{23}) &= \det \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix} = 3, \\ \det(\mathbb{A}_{31}) &= \det \begin{pmatrix} 0 & 1 \\ 2 & 0 \end{pmatrix} = -2, & \det(\mathbb{A}_{32}) &= \det \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} = 0, & \det(\mathbb{A}_{33}) &= \det \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} = 2, \end{aligned}$$

donc on peut calculer  $\det(\mathbb{A})$  par l'une des formules suivantes :

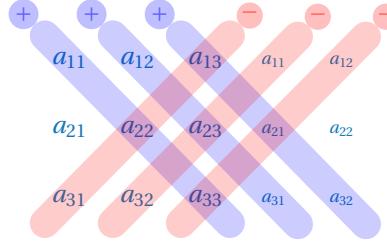
- ★  $1\det(\mathbb{A}_{11}) + 0\det(\mathbb{A}_{12}) + 1\det(\mathbb{A}_{13}) = 10 + 0 + 0 = 10$
- ★  $0\det(\mathbb{A}_{21}) + 2\det(\mathbb{A}_{22}) + 0\det(\mathbb{A}_{23}) = 0 + 2 \times 5 + 0 = 10 \leftarrow \text{formule pratique car il n'y a qu'un déterminant à calculer}$
- ★  $0\det(\mathbb{A}_{31}) + 3\det(\mathbb{A}_{32}) + 5\det(\mathbb{A}_{33}) = 0 + 0 + 5 \times 2 = 10$
- ★  $1\det(\mathbb{A}_{11}) + 0\det(\mathbb{A}_{21}) + 0\det(\mathbb{A}_{31}) = 10 + 0 + 0 = 10 \leftarrow \text{formule pratique car il n'y a qu'un déterminant à calculer}$
- ★  $0\det(\mathbb{A}_{12}) + 2\det(\mathbb{A}_{22}) + 3\det(\mathbb{A}_{32}) = 0 + 2 \times 5 + 0 = 10$
- ★  $1\det(\mathbb{A}_{13}) + 0\det(\mathbb{A}_{23}) + 5\det(\mathbb{A}_{33}) = 0 + 0 + 5 \times 2 = 10$

On peut sinon faire apparaître encore plus de zéros dans la matrice jusqu'à obtenir une matrice triangulaire :

$$\det(\mathbb{A}) = \det \begin{pmatrix} 1 & 0 & 1 \\ 0 & 2 & 0 \\ 0 & 3 & 5 \end{pmatrix} \xrightarrow{L_3 \leftarrow L_3 - \frac{3}{2}L_2} \det \begin{pmatrix} 1 & 0 & 1 \\ 0 & 2 & 0 \\ 0 & 0 & 5 \end{pmatrix} = 10.$$

 **Déterminant d'une matrice d'ordre 3 — méthode pratique (règle de SARRUS)** Soit  $\mathbb{A}$  une matrice carrée d'ordre  $n = 3$ . Alors

$$\det(\mathbb{A}) = \det \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = (a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32}) - (a_{13}a_{22}a_{31} + a_{11}a_{23}a_{32} + a_{12}a_{21}a_{33})$$

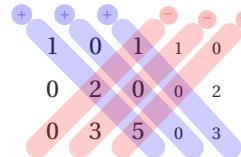


#### EXEMPLE

Soit la matrice

$$\mathbb{A} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 2 & 0 \\ 0 & 3 & 5 \end{pmatrix}$$

alors avec la règle de SARRUS



$$\det(\mathbb{A}) = (1 \times 2 \times 5 + 0 \times 0 \times 0 + 1 \times 0 \times 3) - (1 \times 2 \times 0 + 0 \times 0 \times 3 + 0 \times 0 \times 5) = 10.$$

Si on utilise la définition (règle de LAPLACE), en développant selon la première colonne on obtient

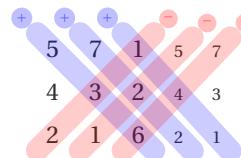
$$\det(\mathbb{A}) = 1 \times \det \begin{pmatrix} 2 & 0 \\ 3 & 5 \end{pmatrix} = 2 \times 5 - 0 \times 3 = 10.$$

#### EXEMPLE

Soit la matrice

$$\mathbb{A} = \begin{pmatrix} 5 & 7 & 1 \\ 4 & 3 & 2 \\ 2 & 1 & 6 \end{pmatrix}$$

alors



$$\det(\mathbb{A}) = (5 \times 3 \times 6 + 7 \times 2 \times 2 + 1 \times 4 \times 1) - (1 \times 3 \times 2 + 5 \times 2 \times 1 + 7 \times 4 \times 6) = -62.$$

**⚠ ATTENTION**

La règle de SARRUS ne s'applique qu'à des matrices d'ordre 3.

**💡 EXEMPLE**

Soit la matrice d'ordre 4 suivante :

$$\mathbb{A} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 2 & 0 & 1 & 0 \\ 1 & 2 & 0 & 4 \\ 1 & 2 & 3 & 0 \end{pmatrix}$$

Alors

$$\det(\mathbb{A}) = \det(\mathbb{A}_{11}) - \det(\mathbb{A}_{14}) = \det \begin{pmatrix} 0 & 1 & 0 \\ 2 & 0 & 4 \\ 2 & 3 & 0 \end{pmatrix} - \det \begin{pmatrix} 2 & 0 & 1 \\ 1 & 2 & 0 \\ 1 & 2 & 3 \end{pmatrix} = -\det \begin{pmatrix} 2 & 4 \\ 2 & 0 \end{pmatrix} - (12 + 0 + 2 - 2 - 0 - 0) = -(-8) - 12 = -4.$$

Si on essaye de «généraliser» la règle de SARRUS on n'obtient pas le bon résultat :

$$(1 \times 0 \times 0 \times 0 + 0 \times 1 \times 4 \times 1 + 0 \times 0 \times 1 \times 2 + 1 \times 2 \times 2 \times 3) - (1 \times 1 \times 2 \times 1 + 1 \times 0 \times 0 \times 2 + 0 \times 2 \times 4 \times 3 + 0 \times 0 \times 1 \times 0) = 10.$$

**📘 Théorème 1.27**

$\mathbb{A}$  est inversible si et seulement si  $\det(\mathbb{A}) \neq 0$ .

**📘 Propriété 1.28**

- ★  $\det(\mathbb{A}^T) = \det(\mathbb{A})$ ,
- ★  $\det(\mathbb{A}^H) = \overline{\det(\mathbb{A})}$ ,
- ★  $\det(\mathbb{A}^{-1}) = \frac{1}{\det(\mathbb{A})}$ ,
- ★  $\det(\mathbb{A}\mathbb{B}) = \det(\mathbb{A}) \cdot \det(\mathbb{B})$ .

**📘 Définition 1.29 (Rang)**

Le RANG d'une matrice quelconque  $\mathbb{A} \in \mathcal{M}_{m,n}$ , noté  $\text{rg}(\mathbb{A})$ , est égal au plus grand entier  $s$  tel que l'on puisse extraire de  $\mathbb{A}$  une matrice carrée d'ordre  $s$  inversible, c'est-à-dire de déterminant non nul. Il représente le nombre maximum de vecteurs colonnes de  $\mathbb{A}$  linéairement indépendants (ou, ce qui est équivalent, le nombre maximum de vecteurs lignes linéairement indépendants).

**✳️ Remarque**

Soit une matrice  $\mathbb{A} \in \mathcal{M}_{m,n}$ . Alors

$$0 \leq \text{rg}(\mathbb{A}) \leq \min(m, n)$$

et  $\text{rg}(\mathbb{A}) = 0$  si et seulement si tous les éléments de  $\mathbb{A}$  sont nuls.

**💡 EXEMPLE**

Soit  $\mathbb{A}$  la matrice suivante

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

Le rang de  $\mathbb{A}$  est 2 car

- ★  $\mathbb{A}$  est d'ordre  $2 \times 3$  donc  $s \leq \min\{2, 3\}$  donc  $s = 0, 1$  ou  $2$ ;
- ★ il existe au moins un élément de  $\mathbb{A}$  différent de zéro, donc  $s \neq 0$ ;
- ★ comme le déterminant de la sous-matrice composée de la première et de la deuxième colonne est nul, on ne peut pas conclure;
- ★ comme le déterminant de la sous-matrice composée de la première et de la troisième colonne est non nul, alors  $s = 2$ .

 EXEMPLE

Soit  $\mathbb{A}$  la matrice suivante

$$\mathbb{A} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 5 & -1 \\ -1 & 0 & -1 \end{pmatrix}.$$

Le rang de  $\mathbb{A}$  est 2 car

- \*  $\mathbb{A}$  est d'ordre  $3 \times 3$  donc  $s \leq 3$ , i.e.  $s = 0, 1, 2$  ou  $3$  ;
- \* il existe au moins un élément de  $\mathbb{A}$  différent de zéro, donc  $s \neq 0$  ;
- \* le déterminant de  $\mathbb{A}$  est 0 donc  $s \neq 3$  ;
- \* le déterminant de la sous-matrice  $\begin{pmatrix} 1 & 0 \\ 0 & 5 \end{pmatrix}$  est non nul, donc  $s = 2$ .

### Opérations élémentaires sur les matrices

 **Définition 1.30 (Opérations élémentaires sur les lignes d'une matrice)**

Les opérations (ou manipulations) élémentaires sur les lignes d'une matrice  $\mathbb{M} \in \mathcal{M}_{m,n}$  sont

- \* la multiplication d'une ligne  $L_i$  par un scalaire non nul  $\alpha$  :

$$L_i \leftarrow \alpha L_i;$$

- \* l'addition d'un multiple d'une ligne  $\alpha L_j$  à une autre ligne  $L_i$  :

$$L_i \leftarrow L_i + \alpha L_j;$$

- \* l'échange de deux lignes :

$$L_i \leftrightarrow L_j.$$

Ces transformations sont équivalentes à la multiplication à gauche (pré-multiplication) de la matrice  $\mathbb{M} \in \mathcal{M}_{m,n}$  par la matrice inversible obtenue en appliquant à la matrice identité  $\mathbb{I}_m$  la transformation correspondante. Par exemple, la transformation qui échange les premières deux lignes de la matrice  $\mathbb{M} \in \mathcal{M}_{4,3}$  suivante

$$\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \\ p & q & r \end{pmatrix} \xrightarrow{L_1 \leftrightarrow L_2} \begin{pmatrix} d & e & f \\ a & b & c \\ g & h & i \\ p & q & r \end{pmatrix}$$

équivaut à multiplier  $\mathbb{M}$  à gauche par la matrice obtenue en échangeant les premières deux lignes de la matrice identité  $\mathbb{I}_4$  :

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \\ p & q & r \end{pmatrix} = \begin{pmatrix} d & e & f \\ a & b & c \\ g & h & i \\ p & q & r \end{pmatrix}$$

 **Définition 1.31 (Opérations élémentaires sur les colonnes d'une matrice)**

Les opérations élémentaires sur les colonnes d'une matrice  $\mathbb{M} \in \mathcal{M}_{m,n}$  sont

- \* la multiplication d'une colonne  $C_i$  par un scalaire  $\alpha$  non nul :

$$C_i \leftarrow \alpha C_i;$$

- \* l'addition d'un multiple d'une colonne  $\alpha C_j$  à une autre colonne  $C_i$  :

$$C_i \leftarrow C_i + \alpha C_j;$$

- \* l'échange de deux colonnes :

$$C_i \leftrightarrow C_j.$$

Ces transformations sont équivalentes à la multiplication à droite (post-multiplication) de la matrice  $\mathbb{M} \in \mathcal{M}_{m,n}$  par la matrice inversible obtenue en appliquant à la matrice identité  $\mathbb{I}_n$  la transformation correspondante. Par exemple la transformation

qui échange les deux premières colonnes de la matrice  $\mathbb{M}$  précédente s'obtient comme suit :

$$\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \\ p & q & r \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} b & a & c \\ e & d & f \\ h & g & i \\ q & p & r \end{pmatrix}$$

### Définition 1.32 (Matrices ÉQUIVALENTES)

Deux matrices sont dites ÉQUIVALENTES si on peut passer de l'une à l'autre par des opérations élémentaires.

### Théorème 1.33

Deux matrices équivalentes ont le même rang.

## 1.2.2 Produits scalaires et vectoriels et normes

On a très souvent besoin, pour quantifier des erreurs ou mesurer des distances, de calculer la “grandeur” d'un vecteur ou d'une matrice. Nous introduisons pour cela la notion de norme vectorielle et celle de norme matricielle.

### Définition 1.34 ( $p$ -norme ou norme de HÖLDER)

On définit la  $p$ -norme (ou norme de HÖLDER) par

$$\|\mathbf{x}\|_p \stackrel{\text{def}}{=} \left( \sum_{i=1}^n |x_i|^p \right)^{1/p}, \quad \text{pour } 1 \leq p < +\infty$$

où les  $x_i$  sont les composantes du vecteur  $\mathbf{x}$ .

Quand on prend  $p = 2$  on retrouve la définition classique de la norme euclidienne.

### Définition 1.35 (Norme infinie ou norme du maximum)

On définit la norme infinie (ou norme du maximum) par

$$\|\mathbf{x}\|_\infty \stackrel{\text{def}}{=} \max_{1 \leq i \leq n} |x_i|.$$

où les  $x_i$  sont les composantes du vecteur  $\mathbf{x}$ .

TO DO

## 1.3 Fonctions trigonométriques

### Fonctions sinus et cosinus

**Domaine de définition** Elles sont définies dans  $\mathbb{R}$  et à valeurs dans  $[-1, 1]$ .

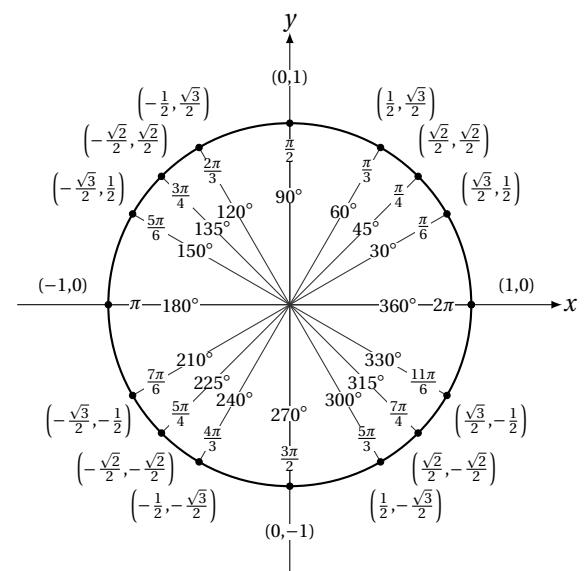
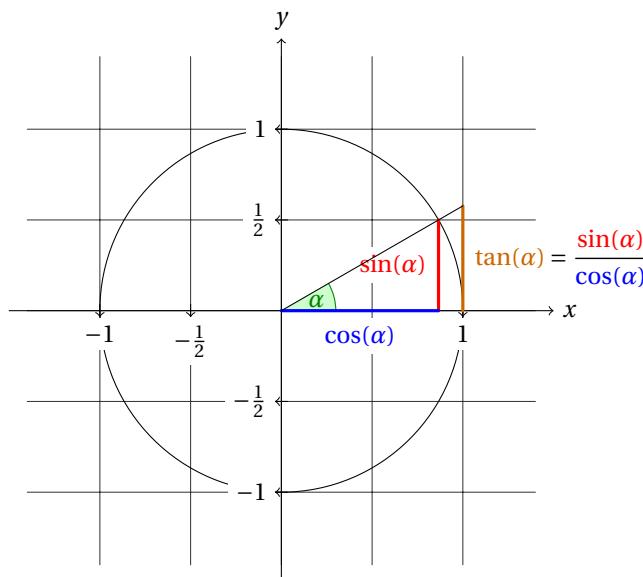
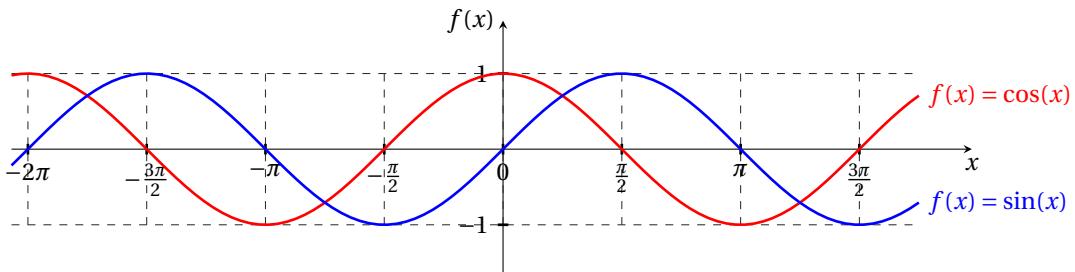
**Périodicité, parité** Elles sont  $2\pi$ -périodiques. La fonction cos est paire, la fonction sin est impaire.

**Dérivées**  $(\cos(x))' = -\sin(x)$ ,  $(\sin(x))' = \cos(x)$ .

*Si  $x \in \mathbb{R}$  est la mesure d'un angle, ces expressions des dérivées ne sont correctes que si  $x$  est exprimé en radians.*

**Limites remarquables**  $\lim_{x \rightarrow 0} \frac{\sin(x)}{x} = 1$ ,  $\lim_{x \rightarrow 0} \frac{1-\cos(x)}{x^2} = \frac{1}{2}$ .

### Représentation graphique



### Fonction tangente

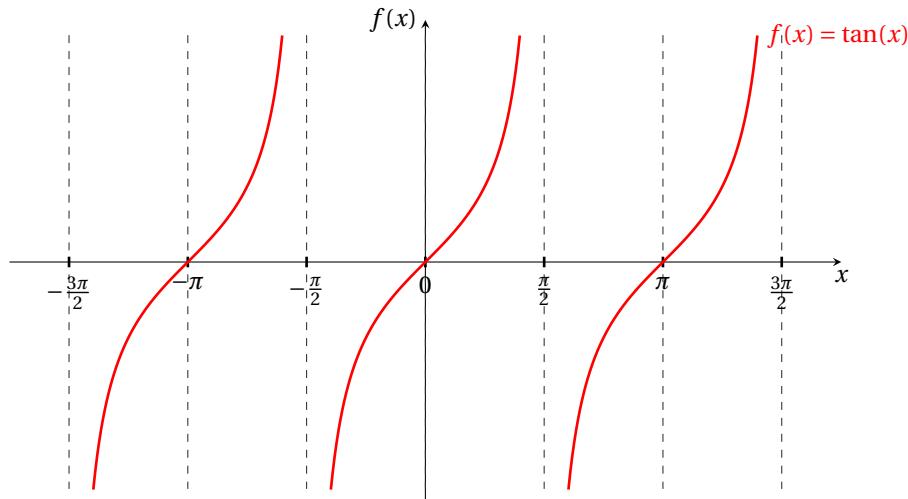
**Domaine de définition** Elle est définie sur  $D := \mathbb{R} \setminus \{\frac{\pi}{2} + k\pi; k \in \mathbb{Z}\}$  par  $\tan(x) := \frac{\sin(x)}{\cos(x)}$ .

**Périodicité, parité** Elle est  $\pi$ -périodique et impaire.

**Dérivée**  $(\tan(x))' = 1 + \tan^2(x) = \frac{1}{\cos^2(x)}$  pour tout  $x \in D$ .

**Limites**  $\lim_{x \rightarrow 0} \frac{\tan(x)}{x} = 1$ .

### Représentation graphique



## Propriétés

$$\begin{array}{lll}
 \cos(\pi - x) = -\cos(x) & \sin(\pi - x) = \sin(x) & \tan(\pi - x) = -\tan(x) \\
 \cos(\pi + x) = -\cos(x) & \sin(\pi + x) = -\sin(x) & \tan(\pi + x) = \tan(x) \\
 \cos(\pi/2 - x) = \sin(x) & \sin(\pi/2 - x) = \cos(x) & \tan(\pi/2 - x) = 1/\tan(x) \\
 \cos(\pi/2 + x) = -\sin(x) & \sin(\pi/2 + x) = \cos(x) & \tan(\pi/2 + x) = -1/\tan(x)
 \end{array}$$

$$\begin{array}{lll}
 \cos(a + b) = \cos(a)\cos(b) - \sin(a)\sin(b) & \cos(a - b) = \cos(a)\cos(b) + \sin(a)\sin(b) \\
 \sin(a + b) = \sin(a)\cos(b) + \cos(a)\sin(b) & \sin(a - b) = \sin(a)\cos(b) - \cos(a)\sin(b) \\
 \tan(a + b) = \frac{\tan(a) + \tan(b)}{1 - \tan(a)\tan(b)} & \tan(a - b) = \frac{\tan(a) - \tan(b)}{1 - \tan(a)\tan(b)}
 \end{array}$$

$$\begin{aligned}
 \cos(2a) &= \cos^2(a) - \sin^2(a) = 2\cos^2(a) - 1 = 1 - 2\sin^2(a) \\
 \sin(2a) &= 2\sin(a)\cos(a) \\
 \tan(2a) &= \frac{2\tan(a)}{1 - \tan^2(a)}
 \end{aligned}$$

$$\begin{array}{ll}
 \cos(a) + \cos(b) = 2\cos\left(\frac{a-b}{2}\right)\cos\left(\frac{a+b}{2}\right) & \cos(a) - \cos(b) = -2\sin\left(\frac{a-b}{2}\right)\sin\left(\frac{a+b}{2}\right) \\
 \sin(a) + \sin(b) = 2\sin\left(\frac{a+b}{2}\right)\cos\left(\frac{a-b}{2}\right) & \sin(a) - \sin(b) = 2\sin\left(\frac{a-b}{2}\right)\cos\left(\frac{a+b}{2}\right)
 \end{array}$$

$$\begin{aligned}
 \cos(a)\cos(b) &= \frac{\cos(a+b) + \cos(a-b)}{2} \\
 \sin(a)\sin(b) &= \frac{\cos(a-b) - \cos(a+b)}{2} \\
 \sin(a)\cos(b) &= \frac{\sin(a+b) + \sin(a-b)}{2}
 \end{aligned}$$

$$\cos^2(a) = \frac{1 + \cos(2a)}{2} \quad \sin^2(a) = \frac{1 - \cos(2a)}{2}$$

$$\begin{aligned}
 \sin(a) &= \pm\sqrt{1 - \cos^2(a)} = \frac{\tan(a)}{\pm\sqrt{1 + \tan^2(a)}}, \\
 \cos(a) &= \pm\sqrt{1 - \sin^2(a)} = \frac{1}{\pm\sqrt{1 + \tan^2(a)}}, \\
 \tan(a) &= \frac{\sin(a)}{\pm\sqrt{1 - \sin^2(a)}} = \frac{\pm\sqrt{1 - \cos^2(a)}}{\cos(a)}.
 \end{aligned}$$

Soit  $t := \tan\left(\frac{a}{2}\right)$ , alors

$$\cos(a) = \frac{1 - t^2}{1 + t^2} \quad \sin(a) = \frac{2t}{1 + t^2} \quad \tan(a) = \frac{2t}{1 - t^2}$$

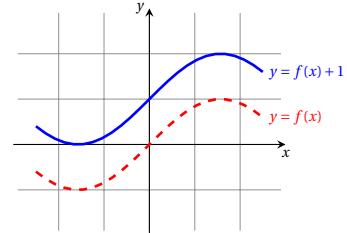
## 1.4 Transformations élémentaires

L'ensemble des points  $(x, y)$  du plan dont les coordonnées satisfont  $y = f(x)$  forme le graphe de la fonction  $f$ . Puisqu'une fonction  $f$  associe une et une seule valeur  $f(x)$  à chacun des points  $x$  de son domaine, une droite verticale coupe le graphe d'une fonction en au plus un point. Lorsque l'on modifie la formule définissant une fonction, le graphe de la fonction se transforme. Certaines modifications de la formule induisent des transformations élémentaires du graphe.

- TRANSLATION VERTICALE

Soit  $g(x) = f(x) + c$ . Lorsque  $c$  est positif, le graphe de  $g$  s'obtient en translatant le graphe de  $f$  vers le haut de  $c$  unités. Lorsque  $c$  est négatif, la translation se fait vers le bas.

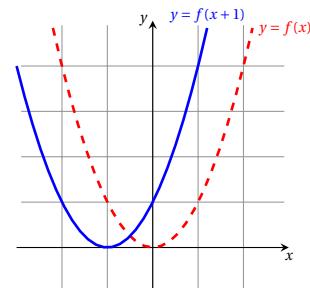
$$\begin{array}{c} g: \mathbb{R} \xrightarrow{f} \mathbb{R} \xrightarrow{t_c} \mathbb{R} \\ x \longmapsto f(x) \longmapsto g(x) = t_c(f(x)) = f(x) + c \end{array}$$



- TRANSLATION HORIZONTALE

Considérons la fonction  $g(x) = f(x+c)$  pour un certain  $c \geq 0$ . La valeur de  $g$  en 0 est égale à la valeur de  $f$  en  $c$ . Le graphe de  $g$  s'obtient en translatant le graphe de  $f$  vers la gauche de  $c$  unités. Lorsque  $c$  est négatif, la translation se fait vers la droite.

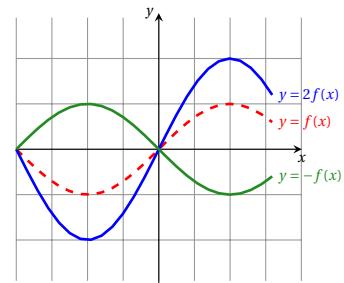
$$\begin{array}{c} g: \mathbb{R} \xrightarrow{t_c} \mathbb{R} \xrightarrow{f} \mathbb{R} \\ x \longmapsto t_c(x) = x+c \longmapsto g(x) = f(t_c(x)) = f(x+c) \end{array}$$



- DILATATION OU CONTRACTION VERTICALE

Soit  $g(x) = cf(x)$  pour un certain  $c$ . La valeur prise par  $g$  en 1 est égale à la valeur prise par  $f$  en 1. Lorsque  $c \geq 1$ , le graphe de  $g$  s'obtient par dilatation du graphe de  $f$  suivant l'axe  $y$  d'un facteur  $c$ . Lorsque  $0 < c \leq 1$ , le graphe de  $g$  s'obtient par contraction de celui de  $f$ . Lorsque  $c$  est négatif, la transformation obtenue est une symétrie par rapport à l'axe  $x$ , suivie d'une dilatation ou d'une contraction d'un facteur  $|c|$ .

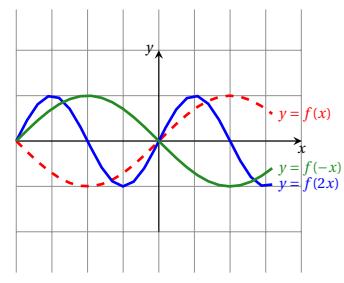
$$\begin{array}{c} g: \mathbb{R} \xrightarrow{t} \mathbb{R} \xrightarrow{d_c} \mathbb{R} \\ x \longmapsto f(x) \longmapsto g(x) = d_c(f(x)) = cf(x) \end{array}$$



- DILATATION OU CONTRACTION HORIZONTALE

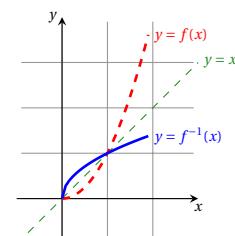
Soit  $g(x) = f(cx)$  pour un certain  $c$ . La valeur prise par  $g$  en 1 est égale à la valeur prise par  $f$  en 1. Lorsque  $c \geq 1$ , le graphe de  $g$  s'obtient par contraction du graphe de  $f$  suivant l'axe  $x$  d'un facteur  $c$ . Lorsque  $0 < c \leq 1$ , le graphe de  $g$  s'obtient par dilatation de celui de  $f$ . Lorsque  $c$  est négatif, la transformation obtenue est une symétrie par rapport à l'axe  $y$ , suivie d'une dilatation ou d'une contraction d'un facteur  $|c|$ .

$$\begin{array}{c} g: \mathbb{R} \xrightarrow{d_c} \mathbb{R} \xrightarrow{f} \mathbb{R} \\ x \longmapsto d_c(x) = cx \longmapsto g(x) = f(d_c(x)) = f(cx) \end{array}$$



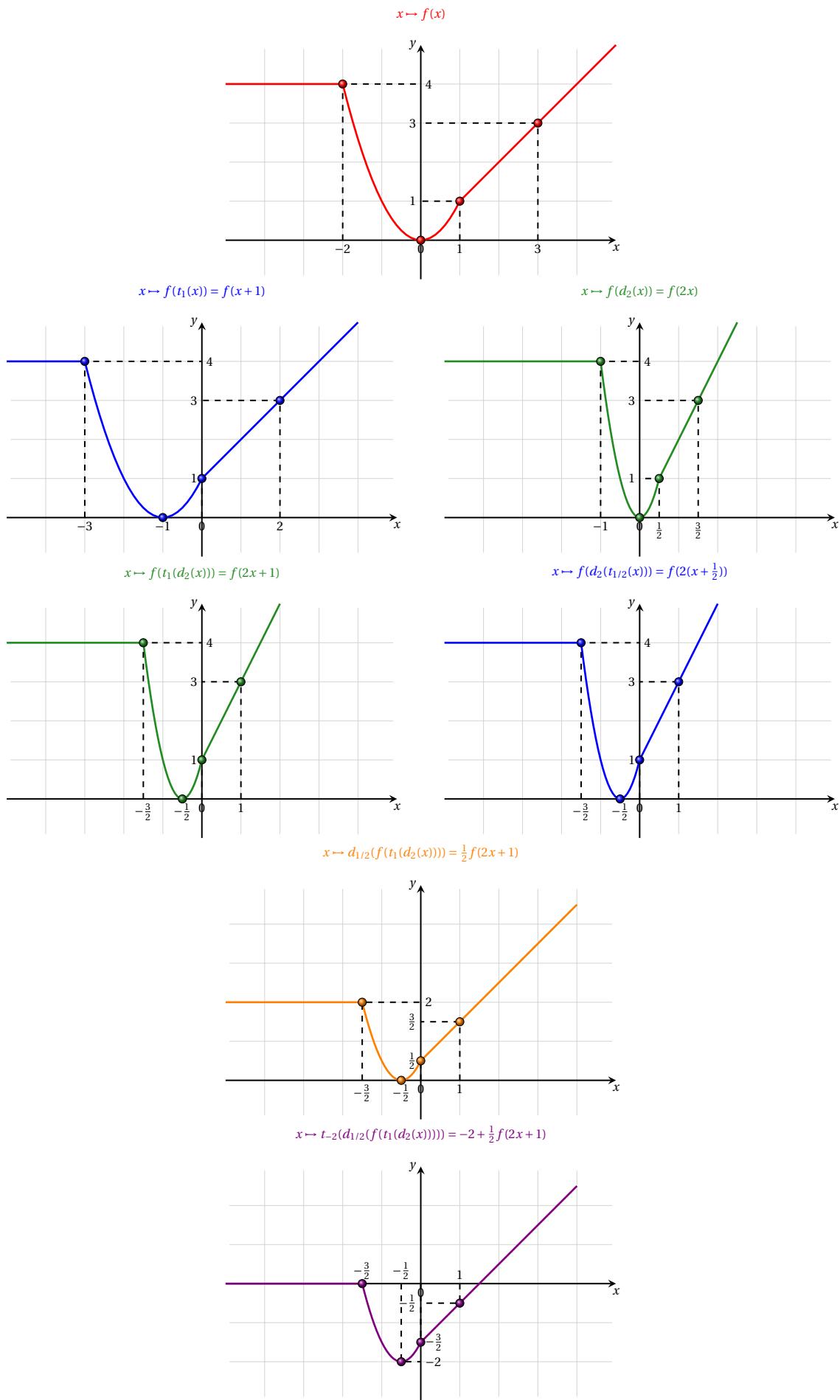
- GRAPHE FONCTION RÉCIPROQUE

Si  $f$  est bijective (donc inversible), le graphe de la réciproque de  $f$  est le symétrique du graphe de  $f$  par rapport à la droite d'équation  $y = x$ .



**Remarque**

Notons que  $t_a(d_b(x)) = (bx) + a$  et  $d_k(t_h(x)) = k(x + h)$  ainsi  $t_a(d_b(x)) = d_k(t_h(x))$  ssi  $b = k$  et  $a = kh$ .



## 1.5 Nombres complexes

**Forme algébrique** Tout nombre  $z \in \mathbb{C}$  s'écrit, de manière unique, sous la forme algébrique  $z = x + iy$  où  $x, y \in \mathbb{R}$  et  $i$  est tel que  $i^2 = -1$ . Le réel  $x$  s'appelle "partie réelle de  $z$ " et se note  $\Re(z)$ . Le réel  $y$  s'appelle "partie imaginaire de  $z$ " et se note  $\Im(z)$ . Si  $y = 0$  alors  $z \in \mathbb{R}$ , si  $x = 0$  alors  $z$  est un imaginaire pur.

**Égalité** Deux nombres complexes  $z = x + iy$  et  $z' = x' + iy'$  sont égaux ssi  $(x, y) = (x', y')$ .

Attention : il n'y a pas d'inégalités en  $\mathbb{C}$ .

**Opérations** Soient deux nombres complexes  $z = x + iy$  et  $z' = x' + iy'$ .

$$z + z' = (x + x') + i(y + y'), \quad zz' = (xx' - yy') + i(xy' + x'y).$$

**Conjugué** Le conjugué d'un nombre complexe  $z = x + iy$  est le nombre complexe  $\bar{z} = x + i(-y)$  et on a

$$\bar{\bar{z}} = z, \quad \overline{z + z'} = \bar{z} + \bar{z'}, \quad \overline{zz'} = \bar{z}\bar{z'}, \quad \overline{\left(\frac{z}{z'}\right)} = \frac{\bar{z}}{\bar{z}'}, \quad z + \bar{z} = 2\Re(z), \quad z - \bar{z} = 2\Im(z).$$

**Module** Le module d'un nombre complexe  $z = x + iy$  est le nombre réel positif  $\sqrt{z\bar{z}} = \sqrt{x^2 + y^2}$ . On le note  $|z|$  ou  $\rho$  ou  $r$ . On a

$$\begin{aligned} |z| = 0 &\iff z = 0; & |\Re(z)| \leq |z|; & |\Im(z)| \leq |z|; \\ |zz'| &= |z| |z'|; & |z^n| &= |z|^n, \text{ pour } n \in \mathbb{N}; & \left|\frac{z}{z'}\right| &= \frac{|z|}{|z'|}; \\ && \left||z| - |z'|\right| &\leq |z - z'| \leq |z| + |z'|. \end{aligned}$$

**Argument** On le note  $\arg(z)$  et il est défini, modulo  $2\pi$ , par  $\cos \vartheta = \frac{x}{\rho}$  et  $\sin \vartheta = \frac{y}{\rho}$ . On a (à  $2k\pi$  près,  $k \in \mathbb{Z}$ )

$$\begin{aligned} \arg(zz') &= \arg(z) + \arg(z'); & \arg(1/z) &= -\arg(z); \\ \arg(z/z') &= \arg(z) - \arg(z'); & \arg(z^n) &= n \arg(z), \text{ pour } n \in \mathbb{Z}. \end{aligned}$$

**Forme trigonométrique** Tout nombre complexe non nul  $z$  s'écrit sous forme trigonométrique  $z = \rho(\cos \vartheta + \sin \vartheta)$  avec  $\rho > 0$  le module de  $z$  et  $\vartheta$  un argument de  $z$ .

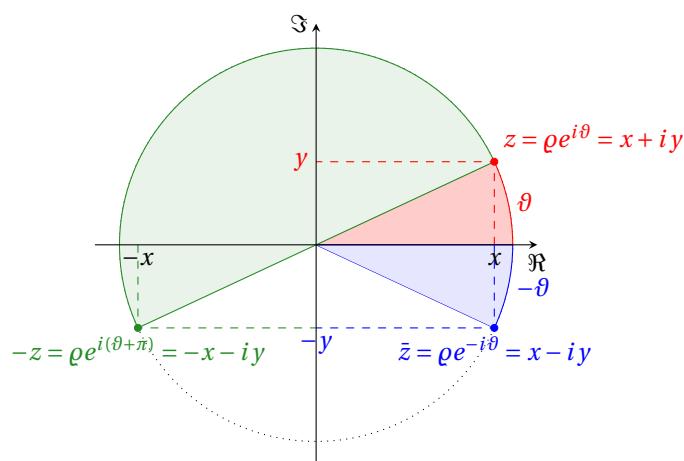
**Forme exponentielle** Tout nombre complexe non nul  $z$  s'écrit sous forme exponentielle  $z = \rho e^{i\vartheta}$ .

\* Formule de Moivre : pour tout  $\vartheta \in \mathbb{R}$  et pour tout  $k \in \mathbb{Z}$

$$z^n = (\rho(\cos \vartheta + \sin \vartheta))^n = \rho^n (\cos(n\vartheta) + \sin(n\vartheta)) = \rho^n e^{in\vartheta}.$$

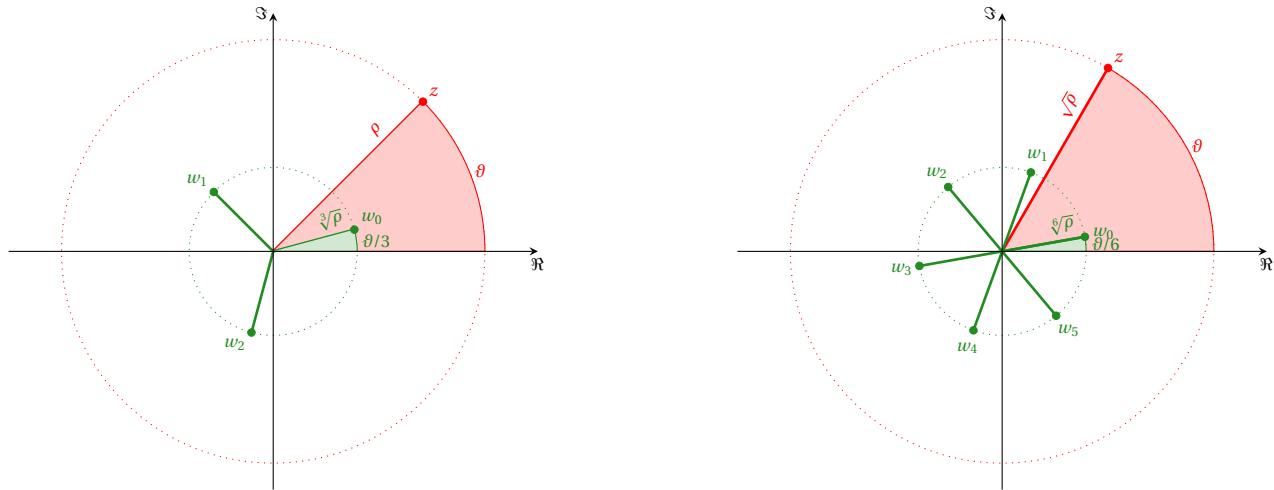
\* Formules d'Euler : pour tout  $x \in \mathbb{R}$  et tout  $k \in \mathbb{Z}$  on a

$$\begin{aligned} \cos x &= \frac{e^{ix} + e^{-ix}}{2}, & \sin x &= \frac{e^{ix} - e^{-ix}}{2i}, \\ \cos(nx) &= \frac{e^{inx} + e^{-inx}}{2}, & \sin(nx) &= \frac{e^{inx} - e^{-inx}}{2i}. \end{aligned}$$



**Racines  $n$ -ièmes** Tout nombre complexe non nul  $z = \rho e^{i\theta}$  possède  $n$  racines  $n$ -ièmes

$$w_k = \sqrt[n]{\rho} e^{i\frac{\theta+2k\pi}{n}} \text{ avec } k \in \{0, 1, \dots, n-1\}.$$



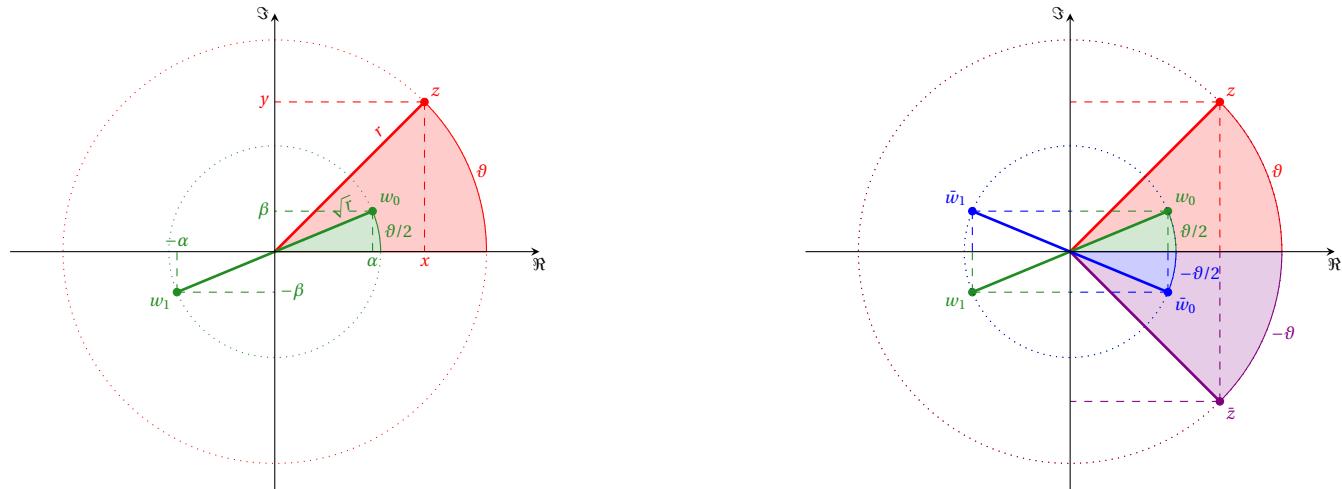
### Astuce (Racines carrées)

Soit  $z = x + iy$  un nombre complexe. Pour déterminer les deux racines carrées  $w_1 = \alpha + i\beta$  et  $w_2 = -w_1$  de  $z$  il est parfois plus simple de procéder par identification, c'est-à-dire de chercher les réels  $\alpha$  et  $\beta$  tels que  $(x + iy) = (\alpha + i\beta)^2$ ; on obtient les relations

$$\alpha^2 - \beta^2 = x, \quad 2\alpha\beta = y, \quad \alpha^2 + \beta^2 = \sqrt{x^2 + y^2}.$$

Autrement dit, en choisissant  $\alpha \geq 0$ ,

$$\alpha = \sqrt{\frac{x + \sqrt{x^2 + y^2}}{2}}, \quad \beta = \operatorname{sgn}(y) \sqrt{\frac{-x + \sqrt{x^2 + y^2}}{2}}.$$



### Proposition 1.36

L'équation  $az^2 + bz + c = 0$ , où  $a \neq 0$ ,  $b$  et  $c$  appartiennent à  $\mathbb{C}$ , admet deux solutions dans  $\mathbb{C}$  (pas nécessairement distinctes) :

$$z_1 = \frac{-b + \delta}{2a}, \quad z_2 = \frac{-b - \delta}{2a},$$

où  $\delta$  et  $-\delta$  sont les deux racines carrées de  $\Delta = b^2 - 4ac$ .

 **Définition 1.37 (Fonctions polynomiales)**

Une application  $P$  de  $\mathbb{C}$  dans  $\mathbb{C}$  est polynomiale s'il existe une famille finie  $\{a_n, \dots, a_1, a_0\}$  de nombres complexes telle que pour tout  $x \in \mathbb{C}$ ,

$$P(x) = a_0 + a_1 x + \cdots + a_n x^n.$$

Pour faire bref, nous ferons un abus de langage en utilisant le terme "polynôme" pour "application polynomiale".

 **Théorème 1.38 (et définition (degré))**

Pour tout polynôme non nul  $P$ , il existe un unique entier  $n \in \mathbb{N}$  et une unique famille  $\{a_n, \dots, a_1, a_0\}$  de nombres complexes telle que

$$a_n \neq 0 \quad \text{et} \quad (\forall x \in \mathbb{C} \quad P(x) = a_0 + a_1 x + \cdots + a_n x^n).$$

L'entier  $n$  est, par définition, le degré de  $P$  (on le note  $d^o P$ ).

 **Remarque**

Le polynôme nul n'a pas de degré, mais par convention  $d^o 0 < d^o g$  pour tout polynôme non nul  $g$ .

 **Théorème 1.39 (Division euclidienne)**

Soient  $f$  et  $g$  deux polynômes avec  $g$  non nul. Alors il existe un unique couple  $(Q, R)$  de polynômes vérifiant

$$\forall x \in \mathbb{C} \quad f(x) = g(x)Q(x) + R(x) \quad \text{et} \quad d^o R < d^o g.$$

$Q$  s'appelle le quotient et  $R$  le reste. On dit que  $f$  est divisible par  $g$  si  $R = 0$ .

 **Définition 1.40 (Racine, multiplicité)**

Un nombre complexe  $a$  est racine d'un polynôme  $P$  si  $P(a) = 0$ . L'ordre de multiplicité d'une racine  $a$  de  $P$  est le plus grand entier  $n$  tel que " $(x - a)^n$ " divise  $P$ .

On montre facilement que  $a$  est racine de  $P$  si et seulement si  $P$  est divisible par  $(x - a)$ .

 **Théorème 1.41 (Décomposition de D'ALEMBERT-GAUSS)**

Soit  $P$  un polynôme de degré supérieur ou égal à 1. Alors l'ensemble des racines de  $P$  est fini et non vide. Notons  $x_1, \dots, x_k$  les  $k$  racines deux à deux distinctes de  $P$  et  $n_i$  l'ordre de multiplicité de  $x_i$ . Alors il existe un unique  $a \in \mathbb{C} \setminus \{0\}$  tel que pour tout  $x \in \mathbb{C}$ ,

$$P(x) = a(x - x_1)^{n_1} \cdots (x - x_k)^{n_k}.$$

On a donc  $n_1 + \cdots + n_k = d^o P$ .

 **Théorème 1.42 (TAYLOR)**

Soit  $P$  un polynôme de degré  $n \geq 1$ . Pour tout  $x_0 \in \mathbb{C}$  et tout  $x \in \mathbb{C}$ ,

$$P(x) = \sum_{k=0}^n P^{(k)}(x_0) \frac{(x - x_0)^k}{k!}$$

où  $P^{(k)}(x_0)$  est la dérivée d'ordre  $k$  en  $x_0$ .

 **Théorème 1.43**

Soient  $P$  un polynôme et  $x_0$  une racine de  $P$ . Elle est de multiplicité  $n$  si et seulement si

$$P^{(n)}(x_0) \neq 0 \quad \text{et} \quad P^{(k)}(x_0) = 0 \text{ pour tout } k \leq n-1.$$

 **Astuce (Recherche d'une racine évidente)**

Les premiers candidats sont toujours 0, 1 et  $-1$ . S'ils ne conviennent pas, au mieux de procéder à l'aveugle, on peut utiliser le fait que, lorsque tous les coefficients  $a_i$  sont des entiers (i.e.  $a_i \in \mathbb{Z}$  pour tout  $i = 0, \dots, n$ ), toute racine «évidente» est du type

$$\frac{\text{diviseur de } a_0}{\text{diviseur de } a_n}.$$

Dès qu'on trouve une racine  $x_0$ , on peut soit arrêter et factoriser le polynôme par  $(x - x_0)$ , soit continuer dans l'espoir de trouver d'autres racines évidentes, ce qui simplifie la factorisation.

## 1.6 Octave/Matlab : guide de survie pour les TP

Nous illustrerons les concepts vu en cours à l'aide de MATLAB (*MATrix LABoratory*), un environnement de programmation et de visualisation. Nous utiliserons aussi GNU Octave (en abrégé Octave) qui est un logiciel libre distribué sous licence GNU GPL. Octave est un interpréteur de haut niveau, compatible la plupart du temps avec MATLAB et possédant la majeure partie de ses fonctionnalités numériques. Dans ce chapitre, nous proposerons une introduction rapide à MATLAB et Octave. Le but de ce chapitre est de fournir suffisamment d'informations pour pouvoir tester les méthodes numériques vues dans ce polycopié. **Il n'est ni un manuel de Octave/Matlab ni une initiation à la programmation.**

### 1.6.1 Les environnements MATLAB et Octave

MATLAB et Octave sont des environnements intégrés pour le Calcul Scientifique et la visualisation. Ils sont écrits principalement en langage C et C++. MATLAB est distribué par la société *The MathWorks* (voir le site [www.mathworks.com](http://www.mathworks.com)). Son nom vient de *MATrix LABoratory*, car il a été initialement développé pour le calcul matriciel. Octave, aussi connu sous le nom de GNU Octave (voir le site [www.octave.org](http://www.octave.org)), est un logiciel distribué gratuitement. Vous pouvez le redistribuer et/ou le modifier selon les termes de la licence GNU *General Public License* (GPL) publiée par la *Free Software Foundation*.

Il existe des différences entre MATLAB et Octave, au niveau des environnements, des langages de programmation ou des *toolbox* (collections de fonctions dédiées à un usage spécifique). Cependant, leur niveau de compatibilité est suffisant pour exécuter la plupart des programmes de ce cours indifféremment avec l'un ou l'autre. Quand ce n'est pas le cas – parce que les commandes n'ont pas la même syntaxe, parce qu'elles fonctionnent différemment ou encore parce qu'elles n'existent pas dans l'un des deux programmes – nous l'indiquons et expliquons comment procéder.

Nous utiliserons souvent dans la suite l'expression “commande MATLAB” : dans ce contexte, MATLAB doit être compris comme le langage utilisé par les deux programmes MATLAB et Octave. De même que MATLAB a ses *toolbox*, Octave possède un vaste ensemble de fonctions disponibles à travers le projet Octave-forge. Ce dépôt de fonctions ne cesse de s'enrichir dans tous les domaines. Certaines fonctions que nous utilisons dans ce polycopié ne font pas partie du noyau d'Octave, toutefois, elles peuvent être téléchargées sur le site [octave.sourceforge.net](http://octave.sourceforge.net). Une fois qu'on a installé MATLAB ou Octave, on peut accéder à l'environnement de travail, caractérisé par le symbole d'invite de commande : `>>` sous MATLAB et `octave:>1` sous Octave. Il représente le prompt : cette marque visuelle indique que le logiciel est prêt à lire une commande. Il suffit de saisir à la suite une instruction puis d'appuyer sur la touche «Entrée».<sup>1</sup>

### 1.6.2 Premiers pas

Lorsqu'on démarre Octave, une nouvelle fenêtre va s'ouvrir, c'est la fenêtre principale qui contient trois onglets : l'onglet “Fenêtre de commandes”, l'onglet “Éditeur” et l'onglet “Documentation”.

L'onglet “Fenêtre de commandes” permet d'entrer directement des commandes et dès qu'on écrit une commande, Octave l'exécute et renvoie instantanément le résultat. L'invite de commande se compose de deux chevrons (`>>`) et représente le prompt : cette marque visuelle indique qu'Octave est prêt à lire une commande. Il suffit de saisir à la suite une instruction puis d'appuyer sur la touche «Entrée». La console Octave fonctionne comme une simple calculatrice : on peut saisir une expression dont la valeur est renvoyée dès qu'on presse la touche «Entrée». Voici un exemple de résolution d'un système d'équations linéaires :

```
>> A = [2 1 0; -1 2 2; 0 1 4]; % Input 3 x 3 matrix
>> b = [1; 2; 3]; % Input column vector
>> soln = A\b % Solve A*x = b by left division
soln =
    0.25000
    0.50000
    0.62500
```

Le symbole `%` indique le début d'un commentaire : tous les caractères entre `%` et la fin de la ligne sont ignorés par l'interpréteur. Le symbole `;` a deux fonctions : il supprime l'affichage d'un résultat intermédiaire et il sépare les lignes d'une matrice. Si on omet le symbole `;`, le résultat sera montré. Par exemple

```
>> A = [2 1 0; -1 2 2; 0 1 4]
A =
    2 1 0
   -1 2 2
    0 1 4
```

Si on ferme Octave et qu'on le relance, comment faire en sorte que l'ordinateur se souvienne de ce que nous avons tapé ? On ne peut pas sauvegarder directement ce qui se trouve dans la onglet “Fenêtre de commandes”, parce que cela comprendrait à la fois les commandes tapées et les réponses du système. Il faut alors avoir un fichier avec uniquement les commandes

1. En attendant l'installation d'Octave sur vos machines, on utilisera la version en ligne <https://octave-online.net/>.

qu'on a tapées et sauver le tout comme un document. Ainsi plus tard on pourra ouvrir ce fichier et lancer Octave sans avoir à retaper toutes les commandes. Passons alors à l'onglet “Éditeur”. Il n'y a rien dans cette nouvelle fenêtre (pas d'en-tête comme dans la “Fenêtre de commandes”). Ce qui veut dire que ce fichier est uniquement pour les commandes : Octave n'interviendra pas avec ses réponses lorsque on écrira le programme et ce tant que on ne le lui demandera pas. Ayant sauvé le programme, pour le faire tourner et afficher les résultats dans la “Fenêtre de commandes” il suffit d'appuyer sur la touche «F5». Si on a fait une faute de frappe, Octave le remarque et demande de corriger. Maintenant qu'on a sauvé le programme, on est capable de le recharger.

### 1.6.3 Notions de base

**Variables et affectation** Dans la plupart des langages informatiques, le nom d'une variable représente une valeur d'un type donné stockée dans un emplacement de mémoire fixe. La valeur peut être modifiée, mais pas le type. Ce n'est pas le cas en Octave, où les variables sont typées dynamiquement. La session interactive suivante illustre ce propos :

```
>> x=1
x = 1
>> x=[2 5]
x =
  2 5

>> x='c'
x = c
```

#### ⚠ ATTENTION

Octave est sensible à la casse. Ainsi, les noms *x* et *X* représentent différents objets. Les noms de variables peuvent être non seulement des lettres, mais aussi des mots ; ils peuvent contenir des chiffres (à condition toutefois de ne pas commencer par un chiffre), ainsi que certains caractères spéciaux comme le tiret bas «\_» (appelé *underscore* en anglais).

<b>ans</b>	Nom pour les résultats
<b>eps</b>	Le plus petit nombre tel que $1+\text{eps} > 1$
<b>inf</b>	$\infty$
<b>NaN</b>	Not a number
<b>i ou j</b>	<i>i</i>
<b>pi</b>	$\pi$

```
>> 5/0
warning: division by zero
ans = Inf
>> 0/0
warning: division by zero
ans = NaN
>> 5*NaN % Most operations with NaN result in NaN
ans = NaN
>> NaN==NaN % Different NaNs are not equal!
ans = 0
>> eps
ans = 2.2204e-16
```

### 1.6.4 Matrices

Pour définir une matrice on doit écrire ses éléments de la première à la dernière ligne, en utilisant le caractère ; pour séparer les lignes (ou aller à la ligne). Par exemple, la commande

```
>> A = [ 1 2 3; 4 5 6]
```

ou la commande

```
>> A = [ 1 2 3; 4 5 6]
```

donnent

```
A =
  1 2 3
  4 5 6
```

c'est-à-dire, une matrice  $2 \times 3$  dont les éléments sont indiqués ci-dessus.

Un vecteur colonne est une matrice  $1 \times n$ , un vecteur ligne est une matrice  $n \times 1$ :

```
>> b = [1 2 3]
b =
  1 2 3

>> b = [1; 2; 3]
b =
  1
  2
  3
```

L'opérateur transposition s'obtient par la commande `'`:

```
>> b = [1 2 3]'
b =
  1
  2
  3
```

Pour extraire les éléments d'une matrice on utilise la commande `A(i, j)` où  $i$  et  $j$  sont la ligne et la colonne respectivement. Une sous-matrice peut être extraite en utilisant le deux points :

```
A(2,3) % element A_{23}
A(:,3) % vecteur colonne [A_{13};...;A_{n3}]
A(1:4,3) % [A_{13};...A_{43}] premières 4 lignes du vecteur colonne [A_{13};...A_{n3}]
A(1,:) % vecteur ligne [A_{11},...,A_{1n}]
A(2,3:end) % [A_{23},...,A_{2n}] vecteur ligne
diag(A) % vecteur colonne [A_{11};...;A_{nn}] contenant la diagonale de A
```

Voici des exemples :

```
>> A = [8 1 6; 3 5 7; 4 9 2]
A =
  8 1 6
  3 5 7
  4 9 2

>> A(2,3) % Element in row 2, column 3
ans = 7
>> A(:,2) % Second column
ans =
  1
  5
  9

>> A(2:3,2:3) % The 2 x 2 submatrix in lower right corner
ans =
  5 7
  9 2
```

### ⚠ ATTENTION

Dans Octave les indices commencent à 1, ainsi `A(1, :)` indique la première ligne, `A(2, :)` la deuxième etc.

La commande `zeros(m, n)` construit la matrice rectangulaire nulle  $\mathbb{O}$ , i.e. celle dont tous les éléments  $a_{ij}$  sont nuls pour  $i = 1, \dots, m$  et  $j = 1, \dots, n$ . La commande `zeros(n)` est un raccourci pour `zeros(n, n)`.

La commande `ones(m, n)` construit une matrice rectangulaire dont les éléments  $a_{ij}$  sont égaux à 1 pour  $i = 1, \dots, m$  et  $j = 1, \dots, n$ . La commande `ones(n)` est un raccourci pour `ones(n, n)`.

La commande `eye(m, n)` renvoie une matrice rectangulaire dont les éléments valent 0 exceptés ceux de la diagonale principale qui valent 1. La commande `eye(n)` (qui est un raccourci pour `eye(n, n)`) renvoie une matrice carrée de dimension  $n$  appelée matrice identité et notée  $\mathbb{I}$ .

Enfin, la commande `A = []` définit une matrice vide.

Construction de vecteurs :

- \* `x=[debut:pas:fin]`
- \* `x=linspace(debut,fin,N)`

```
x = [-5 : 0.25 : 1] % x(k) = -5 + 0.25*(k-1), k=1,2,...,(fin-debut)/pas
y = linspace(-5, 1, 25) % y(k) = -5 + (fin-debut)/N*(k-1), k=1,2,...,N
```

### Opérations arithmétiques

Opérations sur les matrices (lorsque c'est possible) :

- ★ Somme  $C = A + B$ , i.e.  $C_{ij} = A_{ij} + B_{ij}$  :  $C = A + B$
- ★ Produit  $C = AB$ , i.e.  $C_{ij} = \sum_{k=1} A_{ik} + B_{kj}$  :  $C = A * B$
- ★ Division à droite  $C = A \setminus B$  :  $C = A / B$
- ★ Division à gauche  $C = A^{-1}B$  :  $C = A \setminus B$  (si  $B$  est un vecteur colonne,  $C$  est un vecteur colonne solution du système linéaire  $AC = B$ )
- ★ Élévation à la puissance  $C = A \wedge A$  :  $C = A^3$

**Opérations pointées** Quand il s'agit des opérations impliquant des multiplication (donc le produit mais aussi la division et l'élévation à la puissance), la multiplication de deux matrices, avec les notations habituelles, ne signifie pas la multiplication élément par élément mais la multiplication au sens mathématique du produit matriciel. C'est pour cela qu'Octave utilise deux opérateurs distincts pour représenter la multiplication matricielle : `*` et `.`. Le point placé avant l'opérateur indique que l'opération est effectuée élément par élément. Les autres opérations de ce type sont la division à droite et l'élévation à la puissance :

- ★ Produit  $C_{ij} = A_{ij} \cdot B_{ij}$  :  $C = A . * B$
- ★ Division  $C_{ij} = A_{ij} / B_{ij}$  :  $C = A . / B$
- ★ Élévation à la puissance  $C_{ij} = A_{ij}^3$  :  $C = A . ^ 3$

Quand on tente d'effectuer des opérations entre matrices de dimensions incompatibles on obtient un message d'erreur.

**Opérateurs de comparaison** Les opérateurs de comparaison renvoient 1 si la condition est vérifiée, 0 sinon. Ces opérateurs sont

On écrit	Ça signifie
<	<
>	>
$\leq$	$\leq$
$\geq$	$\geq$
$=$	$=$
$\sim =$	$\neq$

### ATTENTION

Les opérateurs de comparaison agissent élément par élément, ainsi lorsqu'on les applique à une matrice le résultat est un matrice qui contient que des 0 ou 1. Par exemple

```
>> A = [1 2 3; 4 -5 6]; B = [7 8 9; 0 1 2];
>> A > B
ans =
  0 0 0
  1 0 1
```

### ATTENTION

Bien distinguer l'instruction d'affectation = du symbole de comparaison ==.

**Connecteurs logiques** Les connecteurs logiques dans Octave sont

On écrit	Ça signifie
<code>&amp;</code>	et
<code> </code>	ou
<code>~</code>	non

On les utilise par exemple pour combiner des conditions complexes (par exemple  $x > -2$  et  $x^2 < 5$ ) :

```
>> (A > B) | (B > 5)
ans =
  1 1 1
  1 0 1
```

## 1.6.5 m-files : Script vs Function

**Script** Un fichier de script contient des instructions qui sont lues et exécutées séquentiellement par l'interpréteur d'Octave. Ce sont obligatoirement des fichiers au format texte. Par exemple, on peut copier les instructions suivantes dans un fichier appelé `sinExp.m` :

```
N = 100;
h = 1/N;
x = 0:h:5;
y = sin(2*pi*x).*exp(-x);
axis([-0.5 1.5 -1.2 1.2])
plot(x, y, x, exp(-x), x, -exp(-x))
grid
```

Ces instructions permettent de tracer un graphe. Si ce fichier se trouve dans le répertoire courant d'Octave, pour l'exécuter il suffit de taper son nom (**sans l'extension**) sur la ligne de commande d'Octave :

```
>> sinExp
```

On peut aussi l'exécuter au moyen de la commande `source` qui prend en argument le nom du fichier ou son chemin d'accès (complet ou relatif au répertoire courant). Par exemple :

```
>> source("Bureau/TP1/sinExp.m")
```

Noter l'usage des points-virgules à la fin de certaines instructions du fichier : ils permettent d'éviter que les résultats de ces instructions soit affiché à l'écran pendant l'exécution du script.

**Function** De très nombreuses fonctions disponibles dans Octave/Matlab sont définies au moyen de la commande `function`. Par convention, chaque définition de fonction est stockée dans un fichier séparé qui porte le nom de la fonction suivi d'une extension `.m` comme pour les fichiers de scripts. Ces fichiers s'appellent des fichiers de fonction. Ils peuvent en réalité contenir plusieurs définitions déclarées au moyen de la commande `function` mais seule la première définition est accessible depuis un script. Les autres définitions concernent des fonctions annexes (on dit parfois des sous-fonctions) qui ne peuvent être utilisées que dans la définition de la fonction principale. À titre d'exemple, écrivons une fonction qui calcule l'aire d'un triangle en fonction des longueurs  $a$ ,  $b$  et  $c$  des côtés grâce à la formule de Héron : Aire =  $\sqrt{p(p - a)(p - b)(p - c)}$  où  $p = (a + b + c)/2$  est le demi-périmètre. On crée pour cela un fichier au format texte appelé `heron.m` contenant les instructions suivantes

```
% Calcule laire s dun triangle par la formule de Heron.
% a, b, c sont les longueurs des artes.
function s = heron(a, b, c)
    p = (a+b+c)/2;
    s = sqrt(p*(p-a)*(p-b)*(p-c));
endfunction
```

La structure type d'un fichier de fonction est la suivante :

- ★ toute ligne commençant par un `#` ou un `%` est considérée comme un commentaire
- ★ les premières lignes du fichier sont des commentaires qui décrivent la syntaxe de la fonction. Ces lignes seront affichées si on utilise la commande `help heron`
- ★ la fonction elle-même est déclarée au moyen de la commande `function` dans laquelle on indique les arguments et la valeur de retour
- ★ cette déclaration est suivie du corps de la définition qui est un bloc d'instructions à exécuter et se termine par le mot-clé `endfunction`

On verra au paragraphe suivant comment faire en sorte que cette fonction soit trouvée automatiquement par Octave. En attendant, la définition donnée ci-dessus peut être testée directement en chargeant le fichier `heron.m` avec la commande `source` et en invoquant la fonction sur la ligne de commande. Par exemple :

```
>> source("Bureau/TP1/heron.m")
>> heron(3,5,4)
ans = 6
```

Pour être vraiment robuste, la définition de la fonction `heron` devrait vérifier que les arguments passés sont bien positifs et que d'autre part ils sont bien tels que chacun est inférieur à la somme des deux autres puisqu'ils sont supposés représenter les longueurs des côtés d'un triangle. Cela revient à dire que le terme sous racine doit être positif. On va compléter la définition en ajoutant une vérification sur les arguments. Pour cela on va définir une sous-fonction appelée `verifArgpas` et une variable `prod` qui donne le résultat du calcul du produit. La fonction `heron` est modifiée comme ceci :

```
% Calcule laire s dun triangle par la formule de Heron.
% a, b, c sont les longueurs des artes.
function s = heron2(a, b, c)
    [ok, prod] = verifArg(a, b, c);
    if (ok == 0),
        error("heron2.m: arguments invalides")
    end;
    s = sqrt(prod);
endfunction

function [ok, prod] = verifArg(a, b, c)
    ok = 0;
    if (a < 0 || b < 0 || c < 0),
        return
    end;
    p = (a+b+c)/2;
    prod = p*(p-a)*(p-b)*(p-c);
    if (prod >= 0),
        ok = 1;
    end;
endfunction
```

Cette fonction doit venir en premier dans le fichier car elle est la fonction principale. La fonction `verifArg` n'est qu'une fonction annexe.

### 1.6.6 Définition, évaluation et graphe de fonctions mathématiques

On se propose d'évaluer en un ensemble de points et tracer la fonction

$$f: [-2; 2] \rightarrow \mathbb{R}$$

$$x \mapsto \frac{1}{1+x^2}$$

#### ⚠ ATTENTION

La variable `x` est un tableau, les opérations `/`, `*` et `\^` agissant sur elle doivent être remplacées par les opérations point correspondantes `./`, `.*` et `\.^` qui opèrent composante par composante.

- ① Dans le fichier `f.m` on écrit la fonction informatique suivante

```
function y=f(x)
    y=1./(1+x.^2);
end
```

Dans un `m-file` ou dans la `prompt` on écrit les instructions suivantes

```
x=0;
y=f(x) % evaluation en un point

x=[-2:0.5:2];

y=f(x) % evaluation en plusieurs points
plot(x,y) % affichage des points (x_i,y_i)

y=feval('f',x) % evaluation en plusieurs points
fplot('f',[-2,2]) % affichage avec evaluation automatique

y=feval(@f,x)
fplot(@f,[-2,2]) % affichage avec evaluation automatique
```

- ② La commande `fplot(fun,lims)` trace le graphe de la fonction `fun` (définie par une chaîne de caractères) sur l'intervalle `]lims(1),lims(2)[`.

```
f='1./(1+x.^2)'
fplot(f,[-2,2])
```

Le graphe est obtenu en échantillonnant la fonction en des abscisses non équiréparties. Il reproduit le graphe réel de  $f(x) = \frac{1}{1+x^2}$  avec une tolérance de 0.2%. Pour améliorer la précision, on pourrait utiliser la commande : `fplot(fun,lims,tol,n,LineSpec)` où `tol` indique la tolérance souhaitée et le paramètre `n` ( $\geq 1$ ) assure que la fonction sera tracée avec un minimum de  $n+1$  points. `LineSpec` spécifie le type de ligne ou la couleur (par exemple, `LineSpec='-'`, pour une ligne en traits discontinus, `LineSpec='r--'` une ligne rouge en traits mixtes, etc.). Pour utiliser les valeurs par défaut de `tol`, `n` ou `LineSpec`, il suffit d'utiliser des matrices vides (`[ ]`).

```
x=0;
y=eval(f)

x=[-2:0.5:2];
y=eval(f);
plot(x,y,'-o')
```

- ③ La commande `inline`, dont la syntaxe usuelle est `fun=inline(expr,arg1, arg2, ..., argn)`, définit une fonction `fun` qui dépend de l'ensemble ordonné de variables `arg1, arg2, ..., argn`. La chaîne de caractères `expr` contient l'expression de `fun`. La forme compacte `fun=inline(expr)` suppose implicitement que `expr` dépend de toutes les variables qui apparaissent dans la définition de la fonction, selon l'ordre alphabétique.

```
f=inline('1./(1+x.^2)')

fplot(f,[-2,2])

x=0;
y=eval(f)
y=f(x)
y=feval(f,x)

x=[-2:0.5:2];
y=f(x);
plot(x,y,'-o')
```

- ④ La syntaxe usuelle d'une fonction anonyme est `fun=@(arg1, arg2,...,argn) [expr]` :

```
f=@(x)[1./(1+x.^2)]

fplot(f,[-2,2])

x=0;
y=f(x)
y=feval(f,x)

x=[-2:0.5:2];
y=f(x);
plot(x,y,'-o')
```

### 1.6.7 Graphes de fonctions

**Tracé de courbes.** Pour tracer le graphe d'une fonction  $f: [a, b] \rightarrow \mathbb{R}$ , Octave a besoin d'une grille de points  $x_i$  où évaluer la fonction, ensuite il relie entre eux les points  $(x_i, f(x_i))$  par des segments. Plus les points sont nombreux, plus le graphe est proche du graphe de la fonction  $f$ . Pour générer les points  $x_i$  on peut utiliser l'instruction `linspace(a,b,n)` qui construit la liste de  $n+1$  éléments

$$\left[ a, a + \frac{b-a}{n}, a + 2\frac{b-a}{n}, \dots, b \right]$$

Voici un exemple avec une sinusoïde :

```
x = linspace(-5,5,101); # x = [-5,-4.9,-4.8,...,5] with 101 elements
y = sin(x); # operation is broadcasted to all elements of the array
plot(x,y)
```

On obtient une courbe sur laquelle on peut zoomer, modifier les marge et sauvegarder dans différents formats (jpg, png, eps...).

On peut même tracer plusieurs courbes sur la même figure. Par exemple, si on veut comparer les graphes de la fonction précédente en modifiant la grille de départ, on peut écrire

-	solid line	--	dashed line
-.	dash-dot line	:	dotted line
.	points	,	pixels
o	circle symbols	^	triangle up symbols
v	triangle down symbols	<	triangle left symbols
>	triangle right symbols	s	square symbols
+	plus symbols	x	cross symbols
D	diamond symbols	*	star symbols
b	blue	g	green
r	red	c	cyan
m	magenta	y	yellow
k	black	w	white

TABLE 1.1 – Quelques options de plot

```
a = linspace(-5,5,5); # a = [-5,-3,-1,1,3,5] with 6 elements
fa = sin(a);
b = linspace(-5,5,10); # b = [-5,-4,-3,...,5] with 11 elements
fb = sin(b);
c = linspace(-5,5,101); # c = [-5,-4.9,-4.8,...,5] with 101 elements
fc = sin(c);
plot(a,fa,b,fb,c,fc)
```

Le résultat est affiché à la figure 1.1a (la courbe bleu correspond à la grille la plus grossière, la courbe rouge correspond à la grille la plus fine).

Pour **tracer plusieurs courbes sur le même graphe**, on peut les mettre les unes à la suite des autres en spécifiant la couleur et le type de trait, changer les étiquettes des axes, donner un titre, ajouter une grille, une légende...

```
x = linspace(-5,5,101); # x = [-5,-4.9,-4.8,...,5] with 101 elements
y1 = sin(x); # operation is broadcasted to all elements of the array
y2 = cos(x);
plot(x,y1,"r-",x,y2,"g.")
legend(['sinus';'cosinus'])
xlabel('abscisses')
ylabel('ordonnees')
title('Comparaison de sin(x) et cos(x)')
grid
```

"r-" indique que la première courbe est à tracer en rouge (red) avec un trait continu, et "g." que la deuxième est à tracer en vert (green) avec des points. Le résultat est affiché à la figure 1.1b. Voir la table 1.1 et la documentation de Matlab pour connaître les autres options.

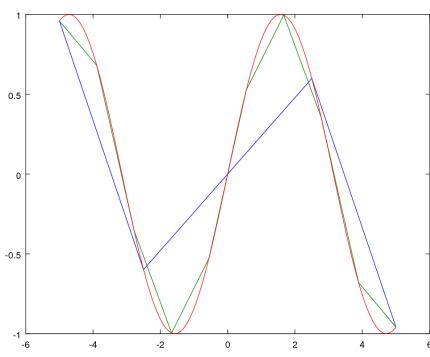
Pour **afficher plusieurs graphes** on a deux possibilités : soit on génère deux fenêtres contenant chacune un graphe, soit on affiche deux graphes côté à côté dans la même figure.

- ★ Un graphe par figure (le résultat est affiché à la figure 1.1c) :

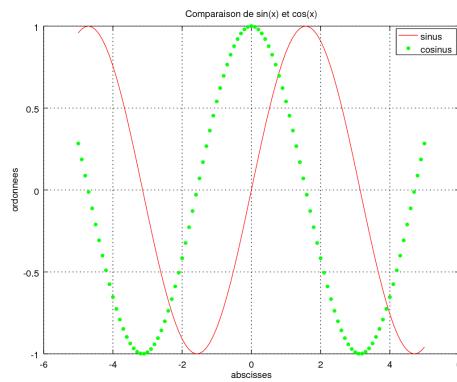
```
x = [-pi:0.05*pi:pi];
figure(1)
plot(x, sin(x), 'r')
figure(2)
plot(x, cos(x), 'g')
```

- ★ Plusieurs graphes côté à côté dans la même figure (le résultat est affiché à la figure 1.1d) : la fonction subplot (xyz) subdivise la figure sous forme d'une matrice (x,y) et chaque case est numérotée, z étant le numéro de la case où afficher le graphe. La numérotation se fait de gauche à droite, puis de haut en bas, en commençant par 1.

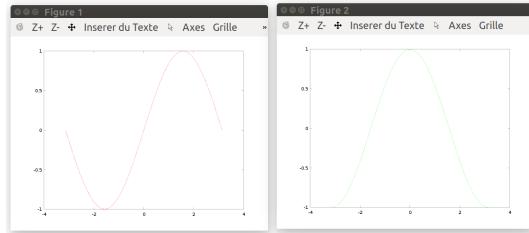
```
x = [-pi:0.05*pi:pi];
subplot(121)
plot(x, sin(x), 'r')
subplot(122)
plot(x, cos(x), 'g')
```



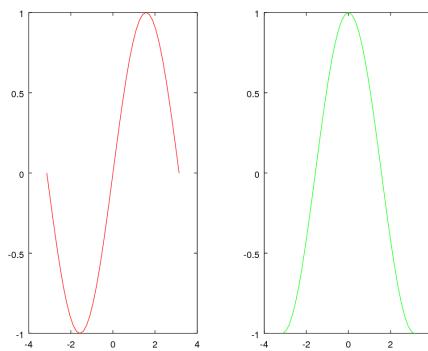
(a) Un graphe avec plusieurs courbes



(b) Un graphe avec plusieurs courbes



(c) Deux figures



(d) Une figure avec deux graphes

FIGURE 1.1 – Exemples pylab

### EXEMPLE

Dans le code suivant on voit comment tracer plusieurs courbes dans un même graphique (avec légende), plusieurs graphes sur une même figure et plusieurs figures.

```
figure(1)
x=[-2:0.5:2];

subplot(2,3,2)
plot(x,x,'r-',x,exp(x),'b*-')
legend(['y=x';'y=e^x'])

subplot(2,3,4)
plot(x,x.^2)
title('y=x^2')

subplot(2,3,5)
plot(x,x.^3)
xlabel('Axe x')

subplot(2,3,6)
plot(x,sqrt(x))

figure(2)
x=linspace(0.1,exp(2),100);
plot(x,log(x));
```

**Tracé d'une fonction bidimensionnelle.** La représentation graphique de l'évolution d'une fonction  $f$  de deux variables  $x$  et  $y$  n'est pas une tâche facile, surtout si le graphique en question est destiné à être imprimé. Dans ce type de cas, un graphe faisant apparaître les lignes de niveaux de la fonction en question peut être une solution intéressante et lisible. Commençons donc par considérer l'exemple simple d'une fonction de ROSENROCK de la forme :

$$f(x, y) = (1 - x)^2 + (y - x^2)^2.$$

Le tracé de cette fonction va nécessiter la création d'un maillage bidimensionnel permettant de stocker l'intervalle de chacune des variables. La fonction destinée à cela s'appelle `meshgrid`. On construit donc le maillage en question sur le rectangle  $[-1; 1] \times [-1; 2]$ . La fonction `meshgrid` fait appel dans ce cas à deux fonctions `linspace` pour chacune des variables.  $Z$  est ici une matrice qui contient les valeurs de la fonction  $f$  sur chaque nœud du maillage. La fonction `surf`???????

```
[x,y] = meshgrid(linspace(-1,1,10),linspace(-1,2,10))
f = @(x,y)(1-x).^2+(y-x).^2.^2;
rangeX = linspace (-1, 1, 10);
rangeY = linspace (-1, 2, 10);
[X, Y] = meshgrid (rangeX, rangeY);
Z = f(X,Y);
surf (X, Y, Z);
```

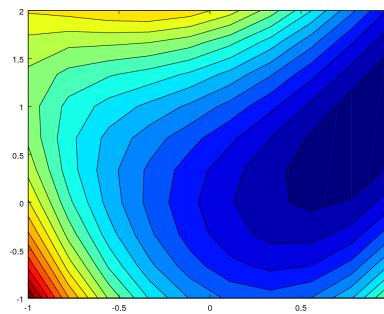
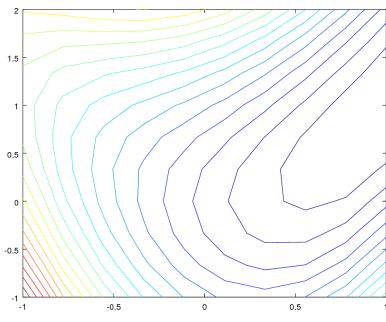
Pour tracer l'évolution de  $f$  en lignes de niveaux on utilisera les fonctions `contour` et `contourf` avec comme arguments les variables  $x$  et  $y$ , les valeurs de  $z$  correspondantes ainsi que le nombre de lignes de niveaux choisis.

La fonction `contour` ne trace que les lignes de niveaux :

```
graphe1 = contour(X,Y,Z,20);
```

La fonction `contourf` colore toute la surface représentant la fonction :

```
graphe3 = contourf(X,Y,Z,20);
colorbar
```



Par défaut, ces courbes de niveaux sont colorées en fonction de leur "altitude"  $z$  correspondante.

### ✿ Remarque (Représentations des erreurs : échelles logarithmique et semi-logarithmique)

Quand on étudie les propriétés de convergence d'une méthode numérique, on trace souvent des graphes représentant

- \* l'erreur  $E$  en fonction de  $h$ , le pas de discrétisation (par exemple pour une formule de quadrature ou le calcul approché de la solution d'une EDO) ;
- \* l'erreur  $E$  en fonction de  $k$ , le pas d'itération (par exemple pour les méthodes de recherche des zéros d'une fonction).

Pour ces graphes on a recours à des représentations en échelle logarithmique ou semi-logarithmique.

- \* Utiliser une **échelle logarithmique** signifie représenter  $\log_{10}(h)$  sur l'axe des abscisses et  $\log_{10}(E)$  sur l'axe des ordonnées. Le but de cette représentation est clair : si  $E = Ch^p$  alors  $\log_{10}(E) = \log_{10}(C) + p\log_{10}(h)$ . En échelle logarithmique,  $p$  représente donc la pente de la ligne droite  $\log_{10}(E)$ . Ainsi, quand on veut comparer deux méthodes, celle présentant la pente la plus forte est celle qui a l'ordre le plus élevé (la pente est  $p = 1$  pour les méthodes d'ordre un,  $p = 2$  pour les méthodes d'ordre deux, et ainsi de suite). Il est très simple d'obtenir avec Python des graphes en échelle logarithmique : il suffit de taper `loglog` au lieu de `plot`. Par exemple, on a tracé sur la figure 1.2a à gauche des droites représentant le comportement de l'erreur de deux méthodes différentes. La ligne rouge correspond à une méthode d'ordre un, la ligne bleue à une méthode d'ordre deux (comparer les deux triangles). Sur la figure 1.2a à droite on a tracé les mêmes données qu'à gauche mais avec la commande `plot`, c'est-à-dire en échelle linéaire pour les axes  $x$  et  $y$ . Il est évident que la représentation linéaire n'est pas la mieux adaptée à ces données puisque la courbe  $E(h) = h^2$  se confond dans ce cas avec l'axe des  $x$  quand  $x \in [10^{-6}; 10^{-2}]$ , bien que l'ordonnée correspondante varie entre  $x \in [10^{-12}; 10^{-4}]$ , c'est-à-dire sur 8 ordres de grandeur.
- \* Plutôt que l'échelle log-log, nous utiliserons parfois une **échelle semi-logarithmique**, c'est-à-dire logarithmique sur l'axe des  $y$  et linéaire sur l'axe des  $x$ . Cette représentation est par exemple préférable quand on trace l'erreur  $E$  d'une méthode itérative en fonction des itérations  $k$  ou plus généralement quand les ordonnées s'étendent sur un intervalle beaucoup plus grand que les abscisses. Si  $E(k) = Ck^n E(0)$  avec  $C \in ]0; 1[$  alors  $\log_{10}(E(k)) = \log_{10}(E(0)) + k^n \log_{10}(C)$ , c'est-à-dire une droite si  $n = 1$ , une parabole si  $n = 2$  etc. La commande pour utiliser l'échelle semi-logarithmique est `semilogy`.

Par exemple, on a tracé sur la figure 1.2b à gauche des courbes représentant le comportement de l'erreur de trois méthodes différentes. La ligne rouge correspond à une méthode d'ordre un, la parabole bleu à une méthode d'ordre

deux et la cubique verte à une méthode d'ordre trois. Sur la figure 1.2b à droite on a tracé les mêmes données qu'à gauche mais avec la commande `plot`, c'est-à-dire en échelle linéaire pour les axes  $x$  et  $y$ . Il est évident que la représentation linéaire n'est pas la mieux adaptée à ces données.

### 1.6.8 Polynômes

Soit  $\mathbb{R}_n[x]$  l'ensemble des polynômes de degré inférieur ou égale à  $n$ ,  $n \in \mathbb{N}$ . Tout polynôme de cet espace vectoriel s'écrit de manière unique comme

$$p_n(x) = \sum_{i=0}^n a_i x^i = a_0 + a_1 x + \cdots + a_n x^n, \quad \text{où } a_i \in \mathbb{R} \text{ pour } i = 0, \dots, n.$$

Les  $n+1$  valeurs réels  $a_0, a_1, \dots, a_n$  sont appelés les **coordonnées de  $p_n$  dans la base canonique**<sup>2</sup> de  $\mathbb{R}_n[x]$  et on peut les stocker dans un vecteur  $\mathbf{p}$  :

$$\mathbf{p} = \text{coord}(p_n, \mathcal{C}_n) = (a_n, a_{n-1}, \dots, a_2, a_1, a_0) \in \mathbb{R}^{n+1}$$

Sous Octave le polynôme  $p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 \in \mathbb{R}_n[x]$  est défini par un vecteur  $\mathbf{p}$  de dimension  $n+1$  contenant les coefficients  $\{a_i\}_{i=0, \dots, n}$  rangés dans l'ordre décroissant des indices, c'est-à-dire que l'on a  $p(1) = a_n, \dots, p(n+1) = a_0$ . Par exemple, pour construire le polynôme  $p(x) = 2 - x + x^2$  nous écrirons

```
p=[1 -1 2]
```

- La commande `polyval` permet d'évaluer le polynôme  $p$  (la fonction polynomiale) en des points donnés. La syntaxe est `polyval(p, x)` où  $x$  est une valeur numérique ou un vecteur. Dans le second cas on obtient un vecteur contenant les valeurs de la fonction polynomiale aux différents points spécifiés dans le vecteur  $\mathbf{x}$ . Par exemple, pour évaluer le polynôme  $p(x) = 1 + 2x + 3x^2$  en  $\mathbf{x} = (-1, 0, 1, 2)$  nous écrirons

```
p=[3 2 1] % p(x)=1+2x+3x^2
y=polyval(p, [-1,0,1,2])
```

- Utilisée avec la commande `fplot`, la commande `polyval` permet de tracer le graphe de la fonction polynomiale sur un intervalle  $[x_{\min}, x_{\max}]$  donné. La syntaxe de l'instruction est `('polyval([a_n, ..., a_0], x)', [x_min, x_max])`. Par exemple, pour tracer le graphe du polynôme  $p(x) = 1 + 2x + 3x^2$  sur l'intervalle  $[-2; 2]$  nous écrirons

```
fplot('polyval([3 2 1],x)', [-2,2])
```

- La commande `roots` calcule les racines du polynôme dans  $\mathbb{C}$ . La syntaxe est `roots(p)`. Par exemple, pour calculer les racines du polynôme  $p(x) = 1 - x^2$  nous écrirons

```
p=[-1 0 1] % p(x)=-x^2+1
racines=roots(p)
```

- La commande `poly` définit un polynôme à partir de ses racines  $r_0, r_1, \dots, r_n$  comme suit :  $p(x) = \prod_{i=0}^n (x - r_i)$ . La syntaxe est `poly(r)` où  $\mathbf{r}$  est un vecteur contenant ses racines. Par exemple, pour définir le polynôme  $p(x) = (x - 1)(x + 1)$  nous écrirons

```
poly([1 -1])
```

- Somme de deux polynômes : si les deux polynômes n'ont pas même degré, il faut ajouter des zéros en début du polynôme de plus petit degré afin de pouvoir calculer l'addition des deux vecteurs représentatifs. Par exemple,

```
p=[1 2 3 4] % p(x)= 4 + 3x + 2x^2 +x^3
q=[0 4 5 6] % q(x)= 6 + 5x + 4x^2 (+0x^3)
s=p+q % s(x)=10 + 8x + 6x^2 +x^3
```

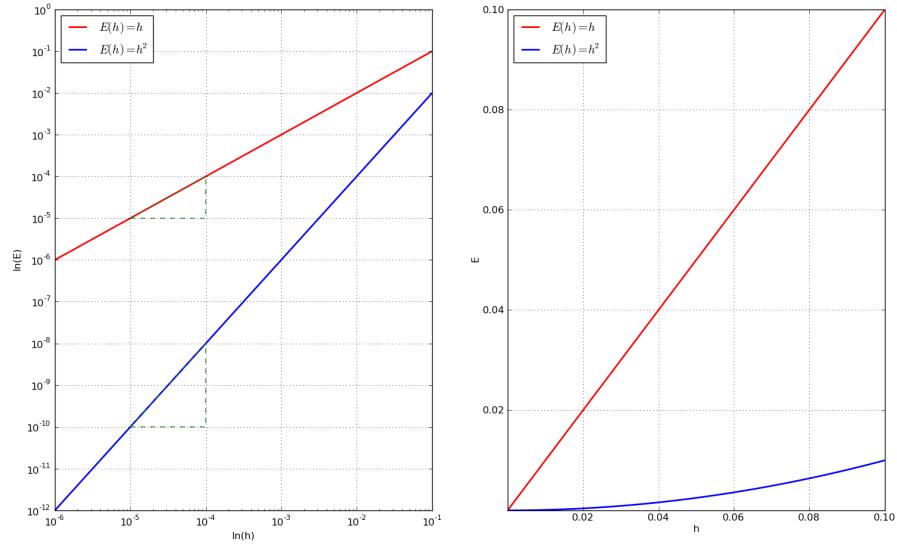
- La commande `conv` permet de calculer le polynôme  $u$  produit de deux polynômes  $p$  et  $q$ . La syntaxe est `conv(p, q)`. Par exemple, pour calculer  $u(x) = p(x)q(x)$  avec  $p(x) = 4 + 3x + 2x^2 + x^3$  et  $q(x) = 6 + 5x + 4x^2$  nous écrirons

```
p=[1 2 3 4] % p(x)=4+3x+2x^2+x^3
q=[4 5 6] % q(x)=6+5x+4x^2
u=conv(p,q) % u(x)=24+38x+43x^2+28x^3+13x^4+4x^5
```

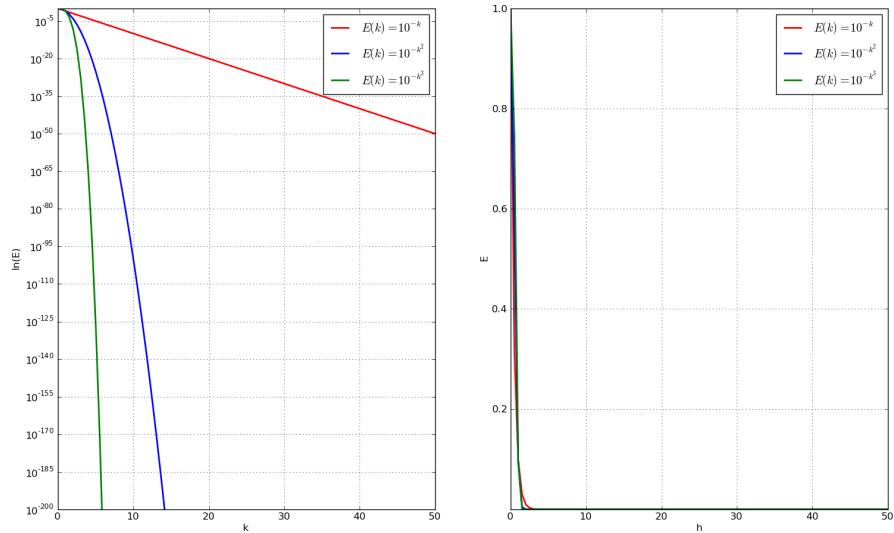
- La commande `deconv` permet de calculer les polynômes  $q$  et  $r$  quotient et reste de la division du polynôme  $u$  par le polynôme  $p$ . La syntaxe est `conv(u, p)`. Par exemple, pour calculer  $q$  et  $r$  tel que  $u(x) = q(x)p(x) + r(x)$  avec  $u(x) = 24 + 38x + 43x^2 + 28x^3 + 13x^4 + 4x^5$  et  $p(x) = 4 + 3x + 2x^2 + x^3$  nous écrirons

---

2. La base canonique de l'espace vectoriel  $\mathbb{R}_n[x]$  est l'ensemble  $\mathcal{C}_n = \{1, x, x^2, \dots, x^n\}$



(a) Échelle logarithmique vs linéaire



(b) Échelle semi-logarithmique vs linéaire

FIGURE 1.2 – Représentations des erreurs.

```
u=[4 13 28 43 38 25] % u(x)=25+38x+43x^2+28x^3+13x^4+4x^5
p=[1 2 3 4] % p(x)=4+3x+3x^2+x^3
[q,r]=deconv(u,p)
```

8. La commande `polyder` permet de calculer le polynôme  $d$  dérivée d'un polynôme  $p$ . La syntaxe est `polyder(p)`. Par exemple, pour calculer  $p'(x)$  avec  $p(x) = 1 + 2x + 3x^2$  nous écrirons

```
p=[3 2 1] % p(x) =1+2x+3x^2
polyder(p) % p'(x)=2+6x
```

9. La commande `polyint` permet de calculer le polynôme  $\int_0^x p(t) dt$  qui s'annule en 0 et qui est une primitive d'un polynôme  $p$ . La syntaxe est `polyint(p)`. Par exemple, pour calculer  $\int_0^x p(t) dt$  avec  $p(x) = 1 + x^2$  nous écrirons

```
p=[1 0 1] % p(x)=1+x^2
integral=polyint(p) % int(p,0..x)=x+x^3/3 donc integral=[1/3 0 1 0]
```

10. Utilisée avec la commande `polyval`, la commande `polyint` permet de calculer l'intégrale d'un polynôme sur un intervalle  $[a, b]$  donné. Par exemple, pour calculer  $\int_0^3 p(t) dt$  avec  $p(x) = 1 + x^2$  nous écrirons

```
area=polyval(integral,3)-polyval(integral,0) % area=3+27/3-0=12
```

11. La commande `polyfit` permet de calculer le polynôme de  $\mathbb{R}_m[x]$  de meilleure approximation au sens des moindres carrés d'un ensemble de points. La syntaxe est `polyfit(xx,yy,m)` où  $xx$  et  $yy$  sont deux vecteurs de  $n$  composantes et  $m$  le degré du polynôme cherché. Si  $m = n$  on obtient le polynôme d'interpolation. Par exemple, pour calculer l'équation de la droite de meilleur approximation de l'ensemble  $\{(0,0.1), (1,0.9), (2,2)\}$  nous écrirons :

```
xx=[0 1 2]
yy=[0.1 0.9 2]
polyfit(xx,yy,1)
```

Pour calculer le polynôme d'interpolation du même ensemble on pose  $m = 2$  et on écrit :

```
polyfit(xx,yy,2)
```

12. Pour afficher le polynôme de façon naturelle il faut utiliser la fonction `polyout`. Par exemple,

```
p=[3 2 1]
polyout(p,'x')
```

## 1.6.9 Structures conditionnelles et répétitives

**La commande if** Supposons vouloir définir la fonction valeur absolue :

$$|x| = \begin{cases} x & \text{si } x \geq 0, \\ -x & \text{sinon.} \end{cases}$$

On a besoin d'une instruction qui opère une disjonction de cas. En Octave il s'agit de l'instruction de choix introduite par le mot-clé `if`. La syntaxe est la suivante :

```
if condition_1
    instruction_1.1
    instruction_1.2
    ...
elseif condition_2
    instruction_2.1
    instruction_2.2
    ...
...
else:
    instruction_n.1
    instruction_n.2
    ...
end
```

où `condition_1`, `condition_2...` représentent des ensembles d'instructions dont la valeur est 1 ou 0 (on les obtient en général en utilisant les opérateurs de comparaison). La première condition `condition_i` ayant la valeur 1 entraîne l'exécution des instructions `instruction_i.1`, `instruction_i.2...`. Si toutes les conditions sont 0, les instructions `instruction_n.1`, `instruction_n.2...` sont exécutées. Les blocs `elseif` et `else` sont optionnels.

Voici un exemple pour établir si un nombre est positif : dans le fichier `sign_of.m` on sauvegarde la fonction

```
function sign=sign_of(a)
if a < 0
    sign = 'negative'
elseif a > 0
    sign = 'positive'
else
    sign = 'zero'
end
```

Dans la “Fenêtre des commandes” on peut alors tester notre fonction :

```
>> sign_of(-1.5)
sign = negative
ans = negative
>> sign_of(2)
sign = positive
ans = positive
>> sign_of(0)
sign = zero
ans = zero
```

**La commande for** La syntaxe de la commande for est schématiquement

```
for var = expression
    instructions
endfor
```

expression peut être un vecteur ou une matrice. Par exemple, le code suivant calcule les 12 premières valeurs de la suite de Fibonacci définie par la relation de récurrence  $u_n = u_{n-1} + u_{n-2}$  avec pour valeurs initiales  $u_0 = u_1 = 1$  :

```
n = 12;
u = ones(1, n);
for i = 3:n
    u(i) = u(i-1)+u(i-2);
end
disp(u)
```

Le résultat affiché est

```
1 1 2 3 5 8 13 21 34 55 89 144
```

**La commande while** La commande while permet de répéter une boucle tant qu'une certaine condition est réalisée. L'exemple suivant cherche à calculer une valeur approchée de la solution  $x$  de l'équation  $x = \cos(x)$ . Le principe est de faire apparaître cette solution comme la limite de la suite  $x_n = \cos(x_{n-1})$  en partant d'une valeur  $x_0$  donnée. On considérera que le résultat est atteint dès que la différence entre deux termes successifs de la suite est inférieure à un seuil de précision donné. Par exemple, si on choisit une précision de 0,001, une valeur initiale de 1 et on impose une limite de 100 itérations afin d'éviter une boucle infinie, le code est le suivant :

```
a = 1;
n = 1;
d = 1;
while d > 0.001 & n < 100
    n = n+1;
    b = cos(a);
    d = abs(b - a);
    a = b;
end
[b, n]
```

Le résultat affiché est :

```
0.73876 18
```

La solution est donc, en arrondissant à trois décimales,  $x = 0,739$  obtenue au bout de 18 itérations.

### ⚠ ATTENTION (LA VECTORISATION, *i.e.* OPTIMISATION DES PERFORMANCES)

La plupart du temps on manipule des vecteurs et des matrices. Les opérateurs et les fonctions élémentaires sont conçus pour favoriser ce type de manipulation et, de manière plus générale, pour permettre la vectorisation des programmes. Certes, le

langage Octave contient des instructions conditionnelles, des boucles et la programmation recursive, mais la vectorisation permet de limiter le recours à ces fonctionnalités qui ne sont jamais très efficaces dans le cas d'un langage interprété. Les surcoûts d'interprétation peuvent être très pénalisants par rapport à ce que ferait un programme C ou FORTRAN compilé lorsque l'on effectue des calculs numériques. Il faut donc veiller à réduire autant que possible le travail d'interprétation en vectorisant les programmes.

#### EXEMPLE

Quasiment toutes les fonctions prédéfinies sont vectorisées.

Le code

```
phi = linspace(0,2*pi,100000);
tic
for i = 1:length(phi),
    sinphi(i) = sin(phi(i));
end;
toc
```

est significativement plus lent que

```
tic
sinphi = sin(phi);
toc
```

#### EXEMPLE

Pour calculer  $\sum_{n=1}^{1000} \frac{1}{n^2}$ , on peut utiliser les trois codes suivants, le deuxième étant significativement plus rapide :

```
n=1:10000;
tic
s=0;
for i = n,
    s+=1/i^2;
end;
s
toc
```

```
tic
s=sum(1./n.^2)
toc
```

```
tic
s=(1./n)*(1./n)';
toc
```

#### EXEMPLE

Pour calculer  $\sum_{i=1}^{100000} x_i y_i$ , on définit deux vecteurs **x** et **y** et on remarque que  $\sum_{i=1}^{100000} x_i y_i = \mathbf{x}^T \mathbf{y}$ .

Le code

```
tic
s = 0;
for i = 1:length(x),
    s+=x(i)*y(i);
end;
toc
```

est significativement plus lent que

```
tic
s=x'*y
toc
```





## Exercices



### ★ Exercice 1.1

Copier les instructions suivantes dans des *script files*. Exécuter les *script* et commenter les résultats.

- ① Somme et produit de matrices, calcul du déterminant et de l'inverse d'une matrice :

```
A=[1 2 3; 4 5 6]
B=[7 8 9; 10 11 12]
C=[13 14; 15 16; 17 18]
A+B
A*C
A+C
inv(A)
det(A)
A=[1 2; 0 0]
inv(A)
```

- ② La commande `diag`

```
v=[2 5 10]
A=diag(v,-1)
v=[2]
A=diag(v,-1)
```

- ③ Matrices triangulaires

```
A =[3 1 2; -1 3 4; -2 -1 3]
L1=tril(A)
L2=tril(A,-1)
```

### Exercice 1.2

Soient les matrices

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

1. Trouver une matrice  $\mathbb{C}$  telle que  $\mathbb{A} - 2\mathbb{B} - \mathbb{C} = \mathbb{O}$ .
2. Trouver une matrice  $\mathbb{D}$  telle que  $\mathbb{A} + \mathbb{B} + \mathbb{C} - 4\mathbb{D} = \mathbb{O}$ .

#### Correction

1. On cherche  $\mathbb{C}$  telle que  $\mathbb{C} = \mathbb{A} - 2\mathbb{B}$ , i.e.

$$\begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \\ c_{31} & c_{32} \end{pmatrix} = \begin{pmatrix} -3 & 2 \\ 0 & 4 \\ 1 & -1 \end{pmatrix} - 2 \begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} -3 - 2 \times 1 & 2 - 2 \times 2 \\ 0 - 2 \times 0 & 4 - 2 \times 1 \\ 1 - 2 \times 1 & -1 - 2 \times 1 \end{pmatrix} = \begin{pmatrix} -5 & -2 \\ 0 & 2 \\ -1 & -3 \end{pmatrix}.$$

2. On cherche  $\mathbb{D}$  telle que  $\mathbb{D} = \frac{1}{4}(\mathbb{A} + \mathbb{B} + \mathbb{C}) = \frac{1}{4}(\mathbb{A} + \mathbb{B} + \mathbb{A} - 2\mathbb{B}) = \frac{1}{2}\mathbb{A} - \frac{1}{4}\mathbb{B}$ , i.e.

$$\begin{pmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \\ d_{31} & d_{32} \end{pmatrix} = \frac{1}{2} \begin{pmatrix} -3 & 2 \\ 0 & 4 \\ 1 & -1 \end{pmatrix} - \frac{1}{4} \begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \times (-3) - \frac{1}{4} \times 1 & \frac{1}{2} \times 2 - \frac{1}{4} \times 2 \\ \frac{1}{2} \times 0 - \frac{1}{4} \times 0 & \frac{1}{2} \times 4 - \frac{1}{4} \times 1 \\ \frac{1}{2} \times 1 - \frac{1}{4} \times 1 & \frac{1}{2} \times (-1) - \frac{1}{4} \times 1 \end{pmatrix} = \begin{pmatrix} -7/4 & 1/2 \\ 0 & 7/4 \\ 1/4 & -3/4 \end{pmatrix}.$$

```
A=[-3 2; 0 4; 1 -1]
B=[1 2; 0 1; 1 1]
C=A-2*B
D=1/4*(A+B+C)
```

### Exercice 1.3

Effectuer les multiplications suivantes

$$\begin{pmatrix} 3 & 1 & 5 \\ 2 & 7 & 0 \end{pmatrix} \begin{pmatrix} 2 & 1 & -1 & 0 \\ 3 & 0 & 1 & 8 \\ 0 & -5 & 3 & 4 \end{pmatrix}, \quad \begin{pmatrix} -3 & 0 & 5 \end{pmatrix} \begin{pmatrix} 2 \\ -4 \\ -3 \end{pmatrix}, \quad \begin{pmatrix} 2 \\ -4 \\ -3 \end{pmatrix} \begin{pmatrix} -3 & 0 & 5 \end{pmatrix}.$$

**Correction**

$$\begin{aligned}
 & \overbrace{\begin{pmatrix} 3 & 1 & 5 \\ 2 & 7 & 0 \end{pmatrix}}^{2 \times 3} \overbrace{\begin{pmatrix} 2 & 1 & -1 & 0 \\ 3 & 0 & 1 & 8 \\ 0 & -5 & 3 & 4 \end{pmatrix}}^{3 \times 4} = \overbrace{\begin{pmatrix} 3 \times 2 + 1 \times 3 + 5 \times 0 & 3 \times 1 + 1 \times 0 + 5 \times (-5) & 3 \times (-1) + 1 \times 1 + 5 \times 3 & 3 \times 0 + 1 \times 8 + 5 \times 4 \\ 2 \times 2 + 7 \times 3 + 0 \times 0 & 2 \times 1 + 7 \times 0 + 0 \times (-5) & 2 \times (-1) + 7 \times 1 + 0 \times 3 & 2 \times 0 + 7 \times 8 + 0 \times 4 \end{pmatrix}}^{2 \times 4} \\
 & = \begin{pmatrix} 9 & -22 & 13 & 28 \\ 25 & 2 & 5 & 56 \end{pmatrix} \\
 & \overbrace{\begin{pmatrix} -3 & 0 & 5 \end{pmatrix}}^{1 \times 3} \overbrace{\begin{pmatrix} 2 \\ -4 \\ -3 \end{pmatrix}}^{3 \times 1} = \overbrace{\begin{pmatrix} -3 \times 2 + 0 \times (-4) + 5 \times (-3) \end{pmatrix}}^{1 \times 1} = -21 \\
 & \overbrace{\begin{pmatrix} 2 \\ -4 \\ -3 \end{pmatrix}}^{3 \times 1} \overbrace{\begin{pmatrix} -3 & 0 & 5 \end{pmatrix}}^{1 \times 3} = \overbrace{\begin{pmatrix} 2 \times (-3) & 2 \times 0 & 2 \times 5 \\ -4 \times (-3) & -4 \times 0 & -4 \times 5 \\ -3 \times (-3) & -3 \times 0 & -3 \times 5 \end{pmatrix}}^{3 \times 3} = \begin{pmatrix} -6 & 0 & 10 \\ 12 & 0 & -20 \\ 9 & 0 & -15 \end{pmatrix}
 \end{aligned}$$

```
[3 1 5; 2 7 0]*[2 1 -1 0; 3 0 1 8; 0 -5 3 4]
[-3 0 5]*[2 -4 -3], % ce qui equivaut a [-3 0 5]*[2; -4; -3]
[2 -4 -3]*[-3 0 5] % ce qui equivaut a [2; -4; -3]*[-3 0 5]
```

**Exercice 1.4**

Calculer  $a, b, c$  et  $d$  tels que

$$\textcircled{1} \quad \begin{pmatrix} 1 & 3 \\ 2 & 8 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \mathbb{I}_2, \quad \textcircled{2} \quad \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 & 3 \\ 2 & 8 \end{pmatrix} = \mathbb{I}_2.$$

Que peut-on conclure ?

**Correction**

Comme

$$\begin{pmatrix} 1 & 3 \\ 2 & 8 \end{pmatrix} \times \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a+3c & b+3d \\ 2a+8c & 2b+8d \end{pmatrix}$$

il faut que

$$\begin{cases} a+3c=1, \\ b+3d=0, \\ 2a+8c=0, \\ 2b+8d=1, \end{cases} \iff \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 4 & -3/2 \\ -1 & 1/2 \end{pmatrix}.$$

De la même manière, pour avoir

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 & 3 \\ 2 & 8 \end{pmatrix} = \begin{pmatrix} a+2b & 3a+8b \\ c+2d & 3c+8d \end{pmatrix}$$

il faut que

$$\begin{cases} a+2b=1, \\ 3a+8b=0, \\ c+2d=0, \\ 3c+8d=1, \end{cases} \iff \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 4 & -3/2 \\ -1 & 1/2 \end{pmatrix}.$$

```
A=[1 3; 2 8]
I2=eye(2)
I2/A
A\I2
```

On conclut que

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 1 & 3 \\ 2 & 8 \end{pmatrix}^{-1} = \begin{pmatrix} 4 & -3/2 \\ -1 & 1/2 \end{pmatrix}.$$

```
inv([1 3; 2 8])
```

### Exercice 1.5

On dit que deux matrices  $\mathbb{A}$  et  $\mathbb{B}$  commutent si  $\mathbb{A}\mathbb{B} = \mathbb{B}\mathbb{A}$ . Trouver toutes les matrices qui commutent avec

$$\mathbb{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 5 \end{pmatrix}.$$

En déduire  $\mathbb{A}^{-1}$ .

#### Correction

On cherche  $\mathbb{B}$  telle que

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 5 \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix} = \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 5 \end{pmatrix}$$

Comme

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 5 \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix} = \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ 3b_{21} & 3b_{22} & 3b_{23} \\ 5b_{31} & 5b_{32} & 5b_{33} \end{pmatrix}$$

et

$$\begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 5 \end{pmatrix} = \begin{pmatrix} b_{11} & 3b_{12} & 5b_{13} \\ b_{21} & 3b_{22} & 5b_{23} \\ b_{31} & 3b_{32} & 5b_{33} \end{pmatrix}$$

il faut que

$$\begin{cases} b_{11} = b_{11}, \\ b_{12} = 3b_{12}, \\ b_{13} = 5b_{13}, \\ 3b_{21} = b_{21}, \\ 3b_{22} = 3b_{22}, \\ 3b_{23} = 5b_{23}, \\ 5b_{31} = b_{31}, \\ 5b_{32} = 3b_{32}, \\ 5b_{33} = 5b_{33}, \end{cases} \iff \mathbb{B} = \begin{pmatrix} \kappa_1 & 0 & 0 \\ 0 & \kappa_2 & 0 \\ 0 & 0 & \kappa_3 \end{pmatrix} \text{ avec } \kappa_1, \kappa_2, \kappa_3 \in \mathbb{R}.$$

Si de plus on veut que  $\mathbb{A}\mathbb{B} = \mathbb{B}\mathbb{A} = \mathbb{I}_3$ , i.e.  $\mathbb{B} = \mathbb{A}^{-1}$ , il faut  $\kappa_1 = 1$ ,  $\kappa_2 = 1/3$  et  $\kappa_3 = 1/5$ .

```
inv(diag([1 3 5]))
```

### Exercice 1.6

Trouver pour quelles valeurs de  $t \in \mathbb{R}$  les matrices suivantes sont inversibles :

$$\mathbb{A} = \begin{pmatrix} t+3 & t^2-9 \\ t^2+9 & t-3 \end{pmatrix}, \quad \mathbb{B} = \begin{pmatrix} t^2-9 & t+3 \\ t-3 & t^2+9 \end{pmatrix}.$$

#### Correction

$$\det(\mathbb{A}) = \det \begin{pmatrix} t+3 & t^2-9 \\ t^2+9 & t-3 \end{pmatrix} = (t+3) \times (t-3) - (t^2-9) \times (t^2+9) = -(t-3)(t+3)(t^2+8).$$

La matrice est inversible pour tout  $t \in \mathbb{R} \setminus \{-3, 3\}$ .

$$\det(\mathbb{B}) = \det \begin{pmatrix} t^2-9 & t+3 \\ t-3 & t^2+9 \end{pmatrix} = (t^2-9) \times (t^2+9) - (t+3) \times (t-3) = (t-3)(t+3)(t^2+8).$$

La matrice est inversible pour tout  $t \in \mathbb{R} \setminus \{-3, 3\}$ .

### Exercice 1.7

Trouver pour quelles valeurs de  $t$  la matrice suivante est inversible

$$\begin{pmatrix} t+3 & -1 & 1 \\ 5 & t-3 & 1 \\ 6 & -6 & t+4 \end{pmatrix}.$$

#### Correction

On commence par calculer le déterminant de la matrice. Étant une matrice d'ordre 3, on peut par exemple utiliser la méthode de SARRUS :

$$\begin{aligned} \det \begin{pmatrix} t+3 & -1 & 1 \\ 5 & t-3 & 1 \\ 6 & -6 & t+4 \end{pmatrix} &= \left( (t+3) \times (t-3) \times (t+4) + 5 \times (-6) \times 1 + 6 \times (-1) \times 5 \right) - \left( 1 \times (t-3) \times 6 + 1 \times (-6) \times (t+3) + (t+4) \times (-1) \times 5 \right) \\ &= t^3 - 4t^2 - 16 = t(t^2 - 4) + 4(t^2 - 4) = (t^2 - 4)(t+4) = (t-2)(t+2)(t+4). \end{aligned}$$

La matrice est inversible pour tout  $t \in \mathbb{R} \setminus \{-4, -2, 2\}$ .

### Exercice 1.8

Soit  $a, b$  et  $c$  trois réels quelconques, calculer les déterminants suivants :

$$D_1 = \det \begin{pmatrix} 1 & 1 & 1 \\ a & b & c \\ a^2 & b^2 & c^2 \end{pmatrix}$$

$$D_2 = \det \begin{pmatrix} 1+a & 1 & 1 \\ 1 & 1+a & 1 \\ 1 & 1 & 1+a \end{pmatrix}$$

#### Correction

Pour calculer un déterminant comportant des paramètres, il est souvent intéressant de faire apparaître des zéros dans une ligne ou une colonne :

$$\begin{aligned} D_1 &= \det \begin{pmatrix} 1 & 1 & 1 \\ a & b & c \\ a^2 & b^2 & c^2 \end{pmatrix} \xrightarrow[C_2 \leftarrow C_2 - C_1 \quad C_3 \leftarrow C_3 - C_1]{=} \det \begin{pmatrix} 1 & 0 & 0 \\ a & b-a & c-a \\ a^2 & b^2-a^2 & c^2-a^2 \end{pmatrix} = \det \begin{pmatrix} b-a & c-a \\ b^2-a^2 & c^2-a^2 \end{pmatrix} \\ &= (b-a)(c^2-a^2) - (c-a)(b^2-a^2) = (b-a)(c-a)((c+a)-(b+a)) = (b-a)(c-a)(c-b); \\ D_2 &= \det \begin{pmatrix} 1+a & 1 & 1 \\ 1 & 1+a & 1 \\ 1 & 1 & 1+a \end{pmatrix} = (1+a)^3 + 1 + 1 - (1+a) - (1+a) - (1+a) = (1+a)^3 - 3(1+a) + 2 \\ &= ((1+a)-1)((1+a)^2 + (1+a) - 2) = ((1+a)-1)((1+a)+2)((1+a)-1) = a^2(3+a). \end{aligned}$$

### Exercice 1.9

1. Pour quelles valeurs de  $\kappa \in \mathbb{R}$  la matrice  $\mathbb{A} = \begin{pmatrix} 1 & \kappa \\ 2 & 3 \end{pmatrix}$  est inversible ?

2. Calculer le rang de la matrice  $\mathbb{B} = \begin{pmatrix} 1 & 2 & 8 \\ 2 & 1 & 4 \\ 0 & 3 & 12 \end{pmatrix}$ .

3. Calculer le rang de la matrice  $\mathbb{C} = \begin{pmatrix} 2 & 1 & 3 \\ 8 & 4 & 12 \\ 1 & 2 & 0 \end{pmatrix}$ .

4. Calculer le déterminant de la matrice  $\mathbb{D} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 2 & 3 & 7 & 4 \\ 3 & 1 & 12 & 0 \\ 4 & 0 & -5 & 0 \end{pmatrix}$ .

5. Calculer le déterminant de la matrice  $\mathbb{E} = \begin{pmatrix} 0 & 2 & 3 & 4 \\ 1 & 7 & 12 & -5 \\ 0 & 3 & 1 & 0 \\ 0 & 4 & 0 & 0 \end{pmatrix}$ .

**Correction**

1. La matrice  $\mathbb{A}$  est inversible pour  $\kappa \neq \frac{3}{2}$  car  $\det(\mathbb{A}) = 3 - 2\kappa$ .

```
determinant=@(k) [det([1 k; 2 3])];
fsolve(determinant,0)
```

2. Sans faire de calcul on peut déjà affirmer que  $1 \leq \text{rg}(\mathbb{B}) \leq 3$ . Comme  $\det(\mathbb{B}) = 0$  (sans faire de calcul, il suffit de remarquer que  $C_3 = 4C_2$ ), alors  $1 \leq \text{rg}(\mathbb{B}) \leq 2$ . Comme  $\det\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} = -3 \neq 0$ , on conclut que  $\text{rg}(\mathbb{B}) = 2$ .

```
rank([1 2 8; 2 1 4; 0 3 12])
```

3. Sans faire de calcul on peut déjà affirmer que  $1 \leq \text{rg}(\mathbb{C}) \leq 3$ . Comme  $\det(\mathbb{C}) = 0$  (sans faire de calcul, il suffit de remarquer que  $L_2 = 4L_1$ ), alors  $1 \leq \text{rg}(\mathbb{C}) \leq 2$ . Comme  $\det\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} = 3 \neq 0$ , on conclut que  $\text{rg}(\mathbb{C}) = 2$ .

```
rank([2 1 3; 8 4 12; 1 2 0])
```

$$4. \det(\mathbb{D}) = \det \begin{pmatrix} 0 & 0 & \boxed{1} & 0 \\ 2 & 3 & 7 & 4 \\ 3 & 1 & 12 & 0 \\ 4 & 0 & -5 & 0 \end{pmatrix} = \det \begin{pmatrix} 2 & 3 & 4 \\ 3 & 1 & 0 \\ \boxed{4} & 0 & 0 \end{pmatrix} = 4 \det \begin{pmatrix} 3 & 4 \\ 1 & 0 \end{pmatrix} = -16.$$

```
det([0 0 1 0; 2 3 7 4; 3 1 12 0; 4 0 -5 0])
```

$$5. \det(\mathbb{E}) = \det \begin{pmatrix} 0 & 2 & 3 & 4 \\ \boxed{1} & 7 & 12 & -5 \\ 0 & 3 & 1 & 0 \\ 0 & 4 & 0 & 0 \end{pmatrix} = -\det \begin{pmatrix} 2 & 3 & 4 \\ 3 & 1 & 0 \\ \boxed{4} & 0 & 0 \end{pmatrix} = -4 \det \begin{pmatrix} 3 & 4 \\ 1 & 0 \end{pmatrix} = 16.$$

```
det([0 2 3 4; 1 7 12 -5; 0 3 1 0; 0 4 0 0])
```

**★ Exercice 1.10 (Opérations élément par élément, produit scalaire, produit vectoriel)**

- ① Définir le vecteur  $x = [\pi/6 \ \pi/4 \ \pi/3]$  et calculer  $s = \sin(x)$  et  $c = \cos(x)$ . En déduire  $\tan(x)$  à l'aide des vecteurs  $s$  et  $c$ .
- ② Copier les instructions suivantes dans un *script file*. Exécuter le *script* et commenter les résultats.

```
x = [1; 2; 3];
y = [4; 5; 6];
v = x.^2
b = sum(x.^2)
s = y.^x
d = dot(x,y)
u = x.*y
p = x.^x
c = cross(x,y)
```

- ③ Calculer la somme des nombres entiers de 1 à 500. Calculer la somme des carrés des nombres entiers de 1 à 500. Calculer la somme des nombres impairs inférieurs ou égaux à 500. Calculer la somme des nombres pairs inférieurs ou égaux à 500.

**Correction**

- ①  $x = [\pi/6 \ \pi/4 \ \pi/3]$   
 $s = \sin(x)$   
 $c = \cos(x)$   
 $t = s ./ c$   
 $\tan(x)$  % on vérifie qu'on a le bon résultat

- ②  $v$  est le vecteur tel que  $v_i = x_i^2$  pour  $i = 1, 2, 3$ ;  $b = v_1 + v_2 + v_3$ .  
 $s$  est le produit scalaire  $y^T x$ .

$u$  est le vecteur tel que  $u_i = x_i y_i$  pour  $i = 1, 2, 3$ .

$p$  est le vecteur tel que  $p_i = (x_i)^{x_i}$  pour  $i = 1, 2, 3$ .

$c$  est le vecteur obtenu par le produit vectoriel de  $x$  et  $y$ .

③  $\sum_{i=1}^n i = \frac{n(n+1)}{2}$  et  $\sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{6}$  :

```
n=500

sum([1:n])
n*(n+1)/2

sum([1:n].^2)
n*(n+1)*(2*n+1)/6

imp=sum(1:2:n)

pair=sum(2:2:n)
```

### ★ Exercice 1.11 (Vectorisation)

Réécrire les codes suivants sans utiliser de boucles :

① 

```
for i = 1:n
    for j = 1:m
        c(i,j) = a(i,j) + b(i,j);
    end
end
```

② 

```
for i = 1:n-1
    a(i) = b(i+1) - b(i);
end
```

③ 

```
for i = 1:n-1
    if (a(i) > 5)
        a(i) = a(i) -20
    end
end
```

④ 

```
n = length (A);
B = zeros (n, 2);
for i = 1:length (A)
    B(i,:) = [A(i+1)-A(i), (A(i+1) + A(i))/2];
end
```

### Correction

① 

```
c=a+b
```

② 

```
a=b(2:n)-b(1:n-1)
```

soit encore

```
a = diff(b);
```

③ 

```
a(a>5) -= 20
```

④ 

```
B = [diff(A) (:), 0.5*(A(1:end-1)+A(2:end)) (:)]
```

### ★ Exercice 1.12 (Sommes, produits et algèbre linéaire pour éliminer les boucles)

Soient  $\mathbf{u}, \mathbf{v}, \mathbf{w}, \mathbf{x}, \mathbf{y}$  des vecteurs ligne de  $\mathbb{R}^n$ . On se propose de calculer

$$\sum_{i=1}^n u_i v_i, \quad \sum_{i=1}^n w_i x_i^2, \quad \sum_{i=1}^n w_i x_i y_i.$$

### Correction

1. Notons que  $\sum_{i=1}^n u_i v_i = \mathbf{u} \cdot \mathbf{v}^T$  donc

```
u = [1 2 3 4 5];
v = [3 6 8 9 10];

% Avec une boucle
y = 0;
for i=1:length(u)
    y = y + u(i)*v(i);
end
y
```

```
% dot notation
y = sum(u.*v)

% produit scalaire
y = u*v,
y = v*u,
y=dot(u,v)
y=dot(v,u)
```

2.  $\sum_{i=1}^n w_i x_i^2 = \mathbf{x} \mathbf{D}_w \mathbf{x}^T$  avec

$$\mathbf{D}_w = \begin{pmatrix} w_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & & w_n \end{pmatrix}$$

donc

```
w = [0.1 0.25 0.12 0.45 0.98];
x = [9 7 11 12 8];

% Avec une boucle
y = 0;
for i = 1:length(w)
    y = y + w(i)*x(i)^2;
end
```

```
y

% dot notation
y = sum(w.*(x.^2))

% algebre lieaire
y = x*diag(w)*x'
```

$$3. \sum_{i=1}^n w_i x_i y_i = \mathbf{x} \mathbf{D}_w \mathbf{y}^T \text{ donc}$$

```
w = [0.1 0.25 0.12 0.45 0.98];
x = [9 7 11 12 8];
y = [2 5 3 8 0];

% Avec une boucle
z = 0;
for i=1:length(w)
    z = z + w(i)*x(i)*y(i);
end
```

```
z

% dot notation
z = sum(w.*x.*y)

% algebre lieaire
z = x*diag(w)*y';
z = y*diag(x)*w';
z = w*diag(y)*x',
```

### ★ Exercice 1.13

- ① Copier les instructions suivantes dans un *script file*. Exécuter le *script* et commenter les résultats.

```
A=[1 2; 4 5];
B=[1 0; 1 1];
A*B
A.*B
A^2
A.^2
A/B
A.\B
```

- ② Afficher la table de multiplication par 1,...,10, *i.e.* la matrice

	1	2	3	4	5	6	7	8	9	10
1	1	2	3	4	5	6	7	8	9	10
2	2	4	6	8	10	12	14	16	18	20
3	3	6	9	12	15	18	21	24	27	30
4	4	8	12	16	20	24	28	32	36	40
5	5	10	15	20	25	30	35	40	45	50
6	6	12	18	24	30	36	42	48	54	60
7	7	14	21	28	35	42	49	56	63	70
8	8	16	24	32	40	48	56	64	72	80
9	9	18	27	36	45	54	63	72	81	90
10	10	20	30	40	50	60	70	80	90	100

### Correction

- ①  $\mathbf{A} * \mathbf{B}$  calcule le produit  $\mathbf{AB} = (\sum_{k=1}^3 a_{ik} b_{kj})_{1 \leq i,j \leq 3}$ ,  $\mathbf{A}.*\mathbf{B}$  calcule la matrice  $\mathbf{C} = (a_{ij} b_{ij})_{1 \leq i,j \leq 3}$ .

```
>> A*B
ans =
3 2
9 5
```

```
>> A.*B
ans =
1 0
4 5
```

$\mathbf{A}^2$  calcule le produit  $\mathbf{AA} = (\sum_{k=1}^3 a_{ik} a_{kj})_{1 \leq i,j \leq 3}$ ,  $\mathbf{A}.^2$  calcule la matrice  $\mathbf{C} = (a_{ij}^2)_{1 \leq i,j \leq 3}$ .

```
>> A^2
ans =
 9 12
24 33
```

```
>> A.^2
ans =
 1 4
16 25
```

A/B calcule le produit  $A \cdot B^{-1}$  si  $B$  est inversible, A.\B calcule la matrice  $C = (a_{ij}/b_{ij})_{1 \leq i,j \leq 3}$ .

```
>> A/B
ans =
 -1 2
 -1 5
```

```
>> A.\B
ans =
 1.00000 0.00000
 0.25000 0.20000
```

- ② Avec les instructions suivantes, on construit la matrice  $A$  qui contient juste les produits, et la matrice  $B$  qui contient aussi l'entête des lignes et colonnes :

```
A(:, :)=[1:10]'.*[1:10]
B(:, :)=[1,1:10]'.*[1,1:10]
```

### ★ Exercice 1.14 (Construction de matrices)

- ① Écrire les instructions pour construire une matrice triangulaire supérieure de dimension 10 ayant des 2 sur la diagonale principale et des -3 sur la seconde sur-diagonale.

$$\begin{bmatrix} 2 & 0 & -3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & -3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & -3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & -3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & -3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 & -3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & -3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & -3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \end{bmatrix}$$

- ② Écrire les instructions permettant d'interchanger la troisième et la septième ligne de la matrice construite au point précédent, puis les instructions permettant d'échanger la quatrième et la huitième colonne.

- ③ Vérifier si les vecteurs suivants de  $\mathbb{R}^4$  sont linéairement indépendants :

```
v1 = [0 1 0 1]
v2 = [1 2 3 4]
v3 = [1 0 1 0]
v4 = [0 0 1 1]
```

- ④ En utilisant la commande `diag`, définir une matrice  $A$  de dimension 10 ayant des 2 sur la diagonale principale et des -1 sur la sur-diagonale et sous-diagonale. Ensuite, en calculer le déterminant, les normes  $\|\cdot\|_1$ ,  $\|\cdot\|_2$  et  $\|\cdot\|_\infty$ , le rayon spectral, les valeurs propres et vecteurs propres. Vérifier enfin que  $V^{-1}AV = D$  où  $D$  est la matrice diagonale qui contient les valeurs propres et  $V$  la matrice dont les colonnes sont les vecteurs propres associés.
- ⑤ Écrire la matrice carrée de taille  $n$  comprenant des  $n$  sur la diagonale principale, des  $n-1$  sur les deux lignes qui l'encadrent, etc.
- ⑥ Écrire la matrice à  $n$  lignes et  $m$  colonnes dont la première colonne ne contient que des 1, la deuxième colonne ne contient que des 2, etc. Écrire la matrice à  $m$  lignes et  $n$  colonnes dont la première ligne ne contient que des 1, la deuxième ligne ne contient que des 2, etc.
- ⑦ Écrire la matrice carrée  $A$  de taille  $2n+1$  comportant des 1 sur la  $(n+1)^{\text{ième}}$  ligne et la  $(n+1)^{\text{ième}}$  colonne et des 0 ailleurs.
- ⑧ Écrire la matrice carrée  $Z$  de taille  $n$  comportant des 1 sur la première et dernière ligne et sur la deuxième diagonale et des 0 ailleurs.
- ⑨ Étant donné une liste de nombres retourner la liste obtenue en écrivant d'abord les termes de rang pair suivis des termes de rang impair.

- ⑩ Écrire la matrice  $(n, n)$  dont les éléments sont  $1, 2, \dots, n^2$  écrits dans l'ordre habituel (sur chaque ligne, de la gauche vers la droite, de la première à la dernière ligne).

### Correction

① `U=2*eye(10)+diag(-3*ones(8,1),2)`

- ② On peut échanger les troisième et septième lignes de la matrice (sans modifier la matrice initiale) avec les instructions :

```
r=[1:10]
r(3)=7
r(7)=3
Ur=U(r,:)
```

Remarquer que le caractère : dans `U(r, :)` fait que toutes les colonnes de `U` sont parcourues dans l'ordre croissant habituel (du premier au dernier terme).

Sinon, si on veut modifier la matrice initiale, on peut utiliser l'instruction :

```
U([3 7],:)=U([7 3],:)
```

Pour échanger les quatrième et huitième colonnes on peut écrire

```
c=[1:10]
c(8)=4
r(4)=8
Uc=U(:,c)
```

- ③ On peut construire la matrice  $A = [v_1; v_2; v_3; v_4]$  et utiliser le fait que les colonnes sont linéairement indépendants si le déterminant de  $A$  est différent de 0, ce qui n'est pas vrai dans notre cas.

```
v1 = [0 1 0 1];
v2 = [1 2 3 4];
v3 = [1 0 1 0];
v4 = [0 0 1 1];
det([v1;v2;v3;v4])
ans = 0
```

- ④ `n=10;`  
`A = 2*diag(ones(1,n))+diag(-1*ones(1,n-1),1)+diag(-1*ones(1,n-1),-1)`  
`detA=det(A)`  
`nrm1=norm(A,1)`  
`nrm2=norm(A,2)`  
`nrminf=norm(A,inf)`  
`X=eig(A);`  
`rho = max(abs(X))`  
`[V,D] = eig(A);`  
`erreur=D-inv(V)*A*V;`  
`norm(erreur)`

- ⑤ En utilisant deux boucles

```
for i=1:n
    for j=1:n
        M(i,j)=n-abs(i-j);
    end
end
```

qu'on peut écrire en version compacte

```
for i=1:n
    M(i,1:n)=n-abs(i-[1:n]);
end
```

En utilisant l'instruction `diag`

```
M=diag(n*ones(1,n));
for i=2:n
    M+=diag((n-i+1)*ones(1,n-i+1),i-1)+diag((n-i+1)*ones(1,n-i+1),1-i);
end
```

```

⑥ A(1:5,:)=[1:10].*ones(5,1)
A'

⑦ n=5;
A=zeros(2*n+1);
A(:,n+1)=1;
A(n+1,:)=1

⑧ n=5;
A=eye(n);
A(1,:)=1;
A(n,:)=1;
A=A(:,n:-1:1)

⑨ liste=rand(1,10)
liste2=[liste([2:2:end]) liste([1:2:end])]

⑩ n=5;
M=[1:n^2];
M=reshape(M,n,n)',
```

### ★ Exercice 1.15 (Construction de matrices, vectorisation, script et fonction)

1. Dans un **fichier zorro.m** écrire une fonction appelée zorro qui prend en entrée un entier  $n \in \mathbb{N}^*$  et renvoi la matrice carrée  $Z$  de taille  $n$  comportant des 1 sur la première et dernière ligne et sur la deuxième diagonale et des 0 ailleurs (sans utiliser de boucles).

Par exemple, pour  $n = 5$ , la commande  $Z=zorro(5)$  devra donner

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

2. Dans un fichier **exercice1.m** écrire un **script** pour tester cette fonction pour  $n = 1, \dots, 5$ .

#### Correction

1. Dans le fichier **zorro.m** on écrit la fonction

```

function A=zorro(n)
A=eye(n);
A(1,:)=1;
A(n,:)=1;
A=A(:,n:-1:1);
end
```

2. Dans le fichier **exercice1.m** on écrit le **script**

```

for n=1:5
Z=zorro(n)
end
```

### ★ Exercice 1.16 (Coût (en temps) d'un produit matrice-vecteur)

Exécuter les instructions suivantes et commenter :

```

n=10000;
step=100;
A=rand(n,n);
v=rand(n,1);
T=[];
sizeA=[];
for k=500:step:n
AA = A(1:k,1:k);
vv = v(1:k);
t = cputime;
b = AA*vv;
tt = cputime - t;
T = [T, tt];
sizeA = [sizeA,k];
end
plot(sizeA,T, 'o')
```

**Correction**

L'instruction `a:step:b` intervenant dans la boucle `for` génère tous les nombres de la forme `a+step*k` où `k` est un entier variant de 0 à `kmax`, où `kmax` est le plus grand entier tel que `a+step*kmax` est plus petit que `b` (dans le cas considéré, `a=500`, `b=10000` et `step=100`). La commande `rand(n,m)` définit une matrice  $n \times m$  dont les éléments sont aléatoires. Enfin, `T` est le vecteur contenant les temps CPU nécessaires à chaque produit matrice-vecteur, et `cputime` renvoie le temps CPU (en secondes) consommé par Octave depuis son lancement. Le temps nécessaire à l'exécution d'un programme est donc la différence entre le temps CPU effectif et celui calculé juste avant l'exécution du programme courant, stocké dans la variable `t`. La commande `plot(sizeA,T,'o')`, montre que le temps CPU augmente comme le carré de  $n$  l'ordre de la matrice.

**★ Exercice 1.17 (Boucle for)**

On considère l'algorithme suivant pour calculer  $\pi$  : on génère  $n$  couples  $\{(x_k, y_k)\}$  de nombres aléatoires dans l'intervalle  $[0, 1]$ , puis on calcule le nombre  $m$  de ceux qui se trouvent dans le premier quart du cercle unité.  $\pi$  est la limite de la suite  $4m/n$ . Écrire un programme pour calculer cette suite et observer comment évolue l'erreur quand  $n$  augmente.

**Correction**

La méthode proposée est une méthode de Monte Carlo. Elle est implémentée dans le programme suivant :

```
%format long
%n=3.e6;
x=[1:10]
for i=x
    n=i*1e4;
    coords=rand(n,2);
    z = coords(:,1).^2+coords(:,2).^2;
    m=sum(z<=1);
    err(i)=abs(pi-4*m/n);
end
err
plot(x,err,'-o')
```

La commande `rand` génère une suite de nombres pseudo-aléatoires. L'instruction `z <= 1` se lit de la manière suivante : on teste si  $z(k) \leq 1$  pour chaque composante du vecteur `z` ; si l'inégalité est satisfaite pour la  $k$ -ème composante de `z` (c'est-à-dire, si le point  $(x(k), y(k))$  appartient à l'intérieur du disque unité) on donne la valeur 1, sinon on lui donne la valeur 0. La commande `sum(z <= 1)` calcule la somme de toutes les composantes de ce vecteur, c'est-à-dire le nombre de points se trouvant à l'intérieur du disque unité. On exécute le programme pour différentes valeurs de  $n$ . Plus  $n$  est grand, meilleure est l'approximation de  $\pi$ . Par exemple, pour  $n = 1000$  on obtient 3.1120, tandis qu'avec  $n = 300000$  on a 3.1406 (naturellement, comme les nombres sont générés aléatoirement, les résultats obtenus pour une même valeur de  $n$  peuvent changer à chaque exécution).

**★ Exercice 1.18 (function)**

Comme  $\pi$  est la somme de la série

$$\pi = \sum_{n \in \mathbb{N}} 16^{-n} \left( \frac{4}{8n+1} - \frac{2}{8n+4} - \frac{1}{8n+5} - \frac{1}{8n+6} \right)$$

on peut calculer une approximation de  $\pi$  en sommant les  $n$  premiers termes, pour  $n$  assez grand. Écrire une fonction Octave pour calculer les sommes partielles de cette série. Pour quelles valeurs de  $n$  obtient-on une approximation de  $\pi$  aussi précise que celle fournie par la variable `pi` ?

**Correction**

Pour répondre à la question on peut utiliser la fonction suivante :

```
function [piapproche]=exercice1_12(N)
piapproche=0;
for n=0:N
    n8=8*n;
    piapproche+=( 4/(n8+1) - 2/(n8+4) - 1/(n8+5) - 1/(n8+6) )*(1/16)^n;
end
return
```

Pour  $n = 10$  on obtient une approximation de  $\pi$  qui coïncide (à la précision Octave) avec la variable interne `pi` d'Octave. Cet algorithme est en effet extrêmement efficace et permet le calcul rapide de centaines de chiffres significatifs de  $\pi$ .

### ★ Exercice 1.19 (function, récursivité)

Écrire une fonction Octave récursive qui calcule le  $n$ -ème élément  $f_n$  de la suite de Fibonacci. Écrire une autre fonction qui calcule  $f_n$  en se basant sur la relation

$$\begin{bmatrix} f_n \\ f_{n-1} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} f_{n-1} \\ f_{n-2} \end{bmatrix}$$

Calculer les temps CPU.

#### Correction

Les fonctions suivantes calculent  $f_n$  en utilisant soit la relation  $f_i = f_{i-1} + f_{i-2}$  (fibrec) soit la relation matricielle (fibmat) :

```
function f=fibrec(n)
    if n == 0
        f = 0;
    elseif n == 1
        f = 1;
    else
        f = fibrec(n-1)+fibrec(n-2);
    end
return

function f=fibmat(n)
    f = [0;1];
    A = [1 1; 1 0];
    f = A^n*f ;
    f = f(1);
return
```

### ★ Exercice 1.20 (function)

Fabriquer une fonction-octave qui calcule le volume  $v$  d'un cylindre de révolution de hauteur  $h$  et dont la base est un disque de rayon  $r$ . Cette fonction doit accepter que  $r$  et  $h$  soient des listes de nombres.

#### Correction

```
function v=cylindre(r,h) % h et r sont 2 matrices-lignes
    v = (r.^2).*h *pi;
return
```

Quand on exécute cette fonction, on obtient un tableau à  $n$  lignes et à  $p$  colonnes qui sont les volumes recherchés. Sur les lignes de ce tableau, on lit les volumes pour  $r$  fixé,  $h$  variant. Sur les colonnes, c'est  $h$  qui est fixé.

### ★ Exercice 1.21 (Coïncidences et anniversaires)

Combien faut-il réunir d'individus dans une salle de classe pour être certain que deux d'entre eux possèdent la même date de naissance ? La réponse est presque évidente, il en faut 366 : même si les 365 premières personnes ont un anniversaire différent, la 366<sup>ème</sup> personne aura forcément une date d'anniversaire commune avec une personne déjà présente.<sup>3</sup>

Maintenant, passons à une question moins évidente : combien faut-il réunir de personnes pour avoir une chance sur deux que deux d'entre elles aient le même anniversaire ? Au lieu de nous intéresser à la probabilité que cet événement se produise, on va plutôt s'intéresser à l'événement inverse : quelle est la probabilité pour que  $n$  personnes n'aient pas d'anniversaire en commun ?

1. si  $n = 1$  la probabilité est 1 (100%) : puisqu'il n'y a qu'une personne dans la salle, il y a 1 chance sur 1 pour qu'elle n'ait pas son anniversaire en commun avec quelqu'un d'autre dans la salle (puisque, fatallement, elle est toute seule dans la salle) ;
2. si  $n = 2$  la probabilité est  $364/365$  (99,73%) : la deuxième personne qui entre dans la salle a 364 chances sur 365 pour qu'elle n'ait pas son anniversaire en commun avec la seule autre personne dans la salle ;
3. si  $n = 3$  la probabilité est  $364/365 \times 363/365$  (99,18%) : la troisième personne qui entre dans la salle a 363 chances sur 365 pour qu'elle n'ait pas son anniversaire en commun avec les deux autres personnes dans la salle mais cela sachant que les deux premiers n'ont pas le même anniversaire non plus, puisque la probabilité pour que les deux premiers n'aient pas d'anniversaire en commun est de  $364/365$ , celle pour que les 3 n'aient pas d'anniversaire commun est donc  $364/365 \times 363/365$  ;
4. si  $n = 4$  la probabilité est  $364/365 \times 363/365 \times 362/365$  (98,36%) et ainsi de suite jusqu'à :

3. On va oublier les années bissextiles et le fait que plus d'enfants naissent neuf mois après le premier de l'an que neuf mois après la Toussaint.

5. si  $n = k$  la probabilité est  $364/365 \times 363/365 \times 362/365 \times \cdots \times (365 - k + 1)/365$

On obtient la formule de récurrence

$$\begin{cases} P_1 = 1, \\ P_{k+1} = P_k \frac{365-k+1}{365}. \end{cases}$$

Tracer un graphe qui affiche la probabilité que deux personnes ont la même date de naissance en fonction du nombre de personnes. Calculer pour quel  $k$  on passe sous la barre des 50%.

Source: <http://eljjdx.canalblog.com/archives/2007/01/14/3691670.html>

### Correction

```
clear all
totale=365;
seuil=50/100;
n=[1:totale];
P(1)=1;
for k=1:totale-1
    P(k+1)=(totale-k+1)*P(k)/totale;
end
nP=1-P;
personnes=sum(nP<seuil);
printf(strcat("On passe la barre de \t", num2str(seuil), " pour k=",num2str(personnes), "\n"))

plot(n,nP,'-',n,seuil*ones(length(n)), '-')
axis([1 totale 0 1])
title(strcat("Seuil=", num2str(seuil), " Personnes=",num2str(personnes) ))
grid
```

Dans un groupe de 23 personnes, il y a plus d'une chance sur deux pour que deux personnes de ce groupe aient leur anniversaire le même jour. Ou, dit autrement, il est plus surprenant de ne pas avoir deux personnes qui ont leur anniversaire le même jour que d'avoir deux personnes qui ont leur anniversaire le même jour (et avec 57, on dépasse les 99% de chances!). On peut s'amuser à adapter les calculs à d'autres problèmes, par exemple on a 61% de chances que parmi 5 personnes prises au hasard, deux ont le même signe astrologique :

```
totale=12;
seuil=61/100;
```

### ★ Exercice 1.22 (Conjecture de Syracuse)

Considérons la suite récurrente

$$\begin{cases} u_1 \in \mathbb{N}^* \text{ donné}, \\ u_{n+1} = \begin{cases} \frac{u_n}{2} & \text{si } n \text{ est pair,} \\ 3u_n + 1 & \text{sinon.} \end{cases} \end{cases}$$

Ce problème est couramment appelé Conjecture de Syracuse (mais aussi problème de Syracuse, algorithme de Hasse, problème de Ulam, problème de Kakutani, conjecture de Collatz, conjecture du  $3n + 1$ ). En faisant des tests numériques on remarque que la suite obtenue tombe toujours sur 1 peu importe l'entier choisi<sup>4</sup> au départ mais personne ne sait encore le démontrer. En 2004, la conjecture a été vérifiée pour tous les nombres inférieurs à  $2^{64}$ .

1. Écrire un script qui, pour une valeur de  $u_1 > 1$  donnée, calcule les valeurs de la suite jusqu'à l'apparition du premier 1.
2. Tracer les valeurs de la suite en fonction de leur position (on appelle cela la trajectoire ou le vol), i.e. les points  $\{(n, u_n)\}_{n=1}^{n=N}$
3. Calculer ensuite le *temps de vol*, i.e. le nombre de terme avant l'apparition du premier 1 ; l'*altitude maximale*, i.e. le plus grand terme de la suite et le *facteur d'expansion*, c'est-à-dire l'altitude maximale divisée par le premier terme.

On peut s'amuser à chercher les valeurs de  $u_1$  donnant le plus grand temps de vol ou la plus grande altitude maximale. Parmi les nombres inférieurs à 100, les champions sont 27 (en altitude maximale) et 97 (en durée de vol).

### Correction

```
clear all
```

```
u(1)=50;
nMAX=10000;
```

4. Dès que  $u_i = 1$  pour un certain  $i$ , la suite devient périodique de valeurs 4, 2, 1

```

n=1;
while (n<nMAX && u(n)~==1)
    if rem(u(n),2)==0
        u(n+1)=u(n)/2;
    else
        u(n+1)=3*u(n)+1;
    end
    n+=1;
end
L=length(u);
M=max(u);
plot(u,'*-')
axis([1 L 1 M])
title(strcat("u(1)=",num2str(u(1)),...
    "\n Temps de vol=",num2str(L-1),...
    "\n Altitude Maximale=",num2str(M),...
    "\n Facteur d'expansion=",num2str(M/u(1))))

```

### ★ Exercice 1.23 (Traitement mathématique des images numériques)

Dans cette exercice nous allons nous intéresser à la manipulation d'images. Nous utiliserons des méthodes basées sur l'algèbre linéaire et l'analyse matricielle.

**Les pixels d'une image :** une image numérique en niveaux de gris (*grayscale* image en anglais) est un tableau de valeurs. Chaque case de ce tableau, qui stocke une valeur, se nomme un pixel. En notant  $n$  le nombre de lignes et  $p$  le nombre de colonnes de l'image, on manipule ainsi un tableau de  $n \times p$  pixels.

La figure ci-dessous montre une visualisation d'un tableau carré avec  $n = p = 512$ , ce qui représente  $512 \times 512 = 2^{18} = 262\,144$  pixels. Les appareils photos numériques peuvent enregistrer des images beaucoup plus grandes, avec plusieurs millions de pixels.

Les valeurs des pixels sont enregistrées dans l'ordinateur ou l'appareil photo numérique sous forme de nombres entiers entre 0 et 255, ce qui fait 256 valeurs possibles pour chaque pixel. La valeur 0 correspond au noir et la valeur 255 correspond au blanc. Les valeurs intermédiaires correspondent à des niveaux de gris allant du noir au blanc.

Pour transformer une image en une matrice il suffit d'indiquer dans notre script :

```

A=imread('lena.jpg');
colormap(gray(256));
A=double(A);
[row,col]=size(A)

```

Octave la transforme en matrice avec la fonction `imread`. On a bien une matrice de taille  $512 \times 512$ . On peut voir Léna<sup>5</sup> avec :

```

colormap(gray(256));
imshow(uint8(A));
% uint8(x) convert x to unsigned 8-bit integer type

```



FIGURE 1.3 – Léna (original)

#### 1. Manipulations élémentaires

- 1.1. En utilisant une manipulation élémentaire de la matrice (sans faire de boucles et sans utiliser de fonctions) obtenir l'image 1.8a
- 1.2. On peut modifier le contraste en “mappant” par une fonction croissante plus rapidement ou plus lentement que la fonction identité  $i: [0;255] \rightarrow [0;255]$ ,  $i(g) = g$ . Par exemple, pour avoir une image plus foncée, on pourra utiliser la fonction

$$\begin{aligned} f: [0;255] &\rightarrow [0;255] \\ g &\mapsto g^3/255^2 \end{aligned}$$

Appliquer cette transformation pour obtenir l'image 1.8b.

5. [http://www.lenna.org/full/l\\_hires.jpg](http://www.lenna.org/full/l_hires.jpg)

1.3. Pour obtenir l'image en négatif de Léna il suffit de prendre le complémentaire par rapport à 255

$$\begin{aligned} f: [0; 255] &\rightarrow [0; 255] \\ g &\mapsto 255 - g \end{aligned}$$

Appliquer cette transformation pour obtenir l'image 1.8c.



FIGURE 1.4 – Manipulations élémentaires

## 2. Résolution

Une matrice de taille  $2^9 \times 2^9$  contient  $2^{18}$  entiers ce qui prend pas mal de place en mémoire. On s'intéresse à des méthodes qui permettent d'être plus économique sans pour cela diminuer la qualité esthétique de l'image. Afin de réduire la place de stockage d'une image, on peut réduire sa résolution, c'est-à-dire diminuer le nombre de pixels. La façon la plus simple d'effectuer cette réduction consiste à supprimer des lignes et des colonnes dans l'image de départ. Les figures suivantes montrent ce que l'on obtient si l'on retient une ligne sur 2 et une colonne sur 2 ce qui donne une matrice  $2^8 \times 2^8$ . Appliquer cette transformation pour obtenir l'une des images suivantes



## 3. Quantification

Une autre façon de réduire la place mémoire nécessaire pour le stockage consiste à utiliser moins de nombres entiers pour chaque valeur. On peut par exemple utiliser uniquement des nombres entiers entre 0 et 3, ce qui donnera une image avec uniquement 4 niveaux de gris. Une telle opération se nomme quantification.

On peut effectuer une conversion de l'image d'origine vers une image avec  $2^{8-k}$  niveaux de valeurs en effectuant les remplacements suivant : tous les valeurs entre 0 et  $2^k$  sont remplacées par la valeur 0, puis tous les valeurs entre  $2^k$  et  $2^{k+1}$  sont remplacées par la valeur  $2^k$  etc. Appliquer cette transformation pour obtenir les images suivantes :

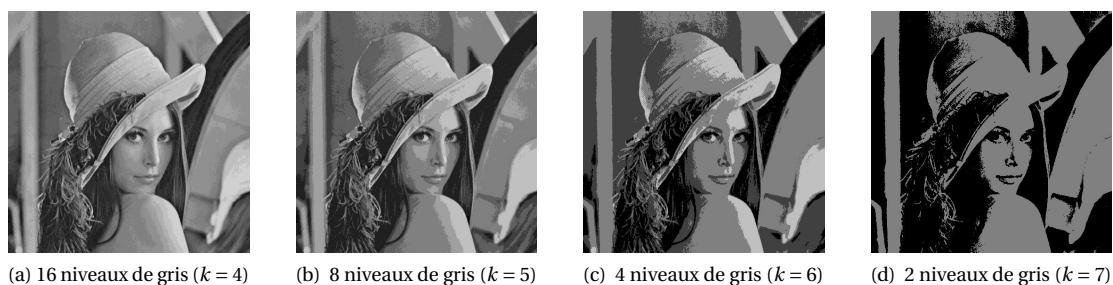


FIGURE 1.5 – Quantification

## 4. Détection des bords

Afin de localiser des objets dans les images, il est nécessaire de détecter les bords de ces objets. Ces bords correspondent à des zones de l'image où les valeurs des pixels changent rapidement. C'est le cas par exemple lorsque l'on passe du chapeau (qui est clair, donc avec des valeurs grandes) à l'arrière plan (qui est sombre, donc avec des valeurs petites).

Afin de savoir si un pixel avec une valeur est le long d'un bord d'un objet, on prend en compte les valeurs de ses quatre voisins (deux horizontalement et deux verticalement). Pour cela nous allons calculer et afficher la norme d'un gradient discret en chaque pixel comme suit :

$$N(A_{i,j}) = \sqrt{(\partial_x A_{i,j})^2 + (\partial_y A_{i,j})^2}$$

avec

$$\begin{aligned} \text{pour } j = 1, \dots, 512, \quad \partial_x A_{i,j} &\simeq \begin{cases} A_{i+1,j} - A_{i,j}, & \text{pour } i = 1 \\ \frac{A_{i+1,j} - A_{i-1,j}}{2}, & \text{pour } i = 2, \dots, 511 \\ A_{i,j} - A_{i-1,j}, & \text{pour } i = 512 \end{cases} \\ \text{pour } i = 1, \dots, 512, \quad \partial_y A_{i,j} &\simeq \begin{cases} A_{i,j+1} - A_{i,j}, & \text{pour } j = 1 \\ \frac{A_{i,j+1} - A_{i,j-1}}{2}, & \text{pour } j = 2, \dots, 511 \\ A_{i,j} - A_{i,j-1}, & \text{pour } j = 512 \end{cases} \end{aligned}$$

Appliquer cette transformation suivie de la transformation en négatif pour obtenir l'image 1.10a. On remarque que les valeurs obtenues appartiennent à l'intervalle [130; 255].

Pour améliorer le rendu, ramener le niveaux de gris à l'intervalle [0; 255] par une transformation affine ce qui donne l'image 1.10b. Cela correspond à la transformation

$$\begin{aligned} f: [m; M] &\rightarrow [0; 255] \\ g &\mapsto \frac{255}{M-m}(x-m) \end{aligned}$$

## 5. Floutage par diffusion

On veut lisser les endroits à fort gradient. Pour cela nous allons calculer une moyenne en chaque pixel comme suit :

$$\mathbb{A} \leftarrow \mathbb{A} + \frac{\partial_{xx}\mathbb{A} + \partial_{yy}\mathbb{A}}{5}$$

avec

$$\begin{aligned} \text{pour } j = 1, \dots, 512, \quad \partial_{xx} A_{i,j} &\simeq \begin{cases} 0, & \text{pour } i = 1 \\ A_{i+1,j} - 2A_{i,j} + A_{i-1,j}, & \text{pour } i = 2, \dots, 511 \\ 0, & \text{pour } i = 512 \end{cases} \\ \text{pour } i = 1, \dots, 512, \quad \partial_{yy} A_{i,j} &\simeq \begin{cases} 0, & \text{pour } j = 1 \\ A_{i,j+1} - 2A_{i,j} + A_{i,j-1}, & \text{pour } j = 2, \dots, 511 \\ 0, & \text{pour } j = 512 \end{cases} \end{aligned}$$

Appliquer 100 fois cette transformation à la matrice  $\mathbb{A}$  pour obtenir l'image 1.10c.<sup>6</sup>

Appliquer ensuite la détection des bords (normalisée) à l'image 1.10c pour obtenir l'image 1.6d.

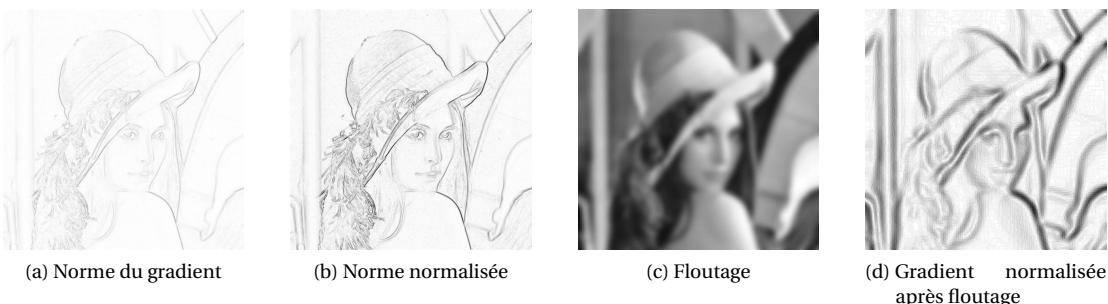


FIGURE 1.6 – Détection des bords et Floutage

## Correction

### 1. Manipulations élémentaires

#### 1.1. Miroir

6. Cela correspond à un schéma 5 points explicite appliquée à l'équation de la chaleur  $\partial_t A = \nabla \cdot (f(\nabla A))$  avec  $f$  l'identité

```

clear all

A=imread('lena512.bmp');
colormap(gray(256));
A=double(A);
[row,col]=size(A)

subplot(1,2,1)

```

```

imshow(uint8(A));
title ( "Original" );
subplot(1,2,2)
B=A(row:-1:1,:);
imshow(uint8(B));
title ( "Miroir" );
imwrite(uint8(B),'exo11.jpg','jpg');

```

## 1.2. Contraste

```

clear all

A=imread('lena512.bmp');
colormap(gray(256));
A=double(A);
[row,col]=size(A)

subplot(1,2,1)

```

```

imshow(uint8(A));
title ( "Original" );
subplot(1,2,2)
B=A.^3/255.^2;
imshow(uint8(B));
title ( "Foncee" );
imwrite(uint8(B),'exo12.jpg','jpg');

```

## 1.3. Négatif

```

clear all

A=imread('lena512.bmp');
colormap(gray(256));
A=double(A);
[row,col]=size(A)

subplot(1,2,1)

```

```

imshow(uint8(A));
title ( "Original" );
subplot(1,2,2)
B=255-A;
imshow(uint8(B));
title ( "Natif" );
imwrite(uint8(B),'exo13.jpg','jpg');

```

## 2. Résolution

```

clear all

A=imread('lena512.bmp');
colormap(gray(256));
A=double(A);
[row,col]=size(A)

subplot(2,2,1)
E=A(1:2:row,1:2:col);
imshow(uint8(E));
imwrite(uint8(E),'exo2E11.jpg','jpg');
subplot(2,2,2)

```

```

E=A(2:2:row,1:2:col);
imwrite(uint8(E),'exo2E21.jpg','jpg');
imshow(uint8(E));
subplot(2,2,3)
E=A(1:2:row,2:2:col);
imshow(uint8(E));
imwrite(uint8(E),'exo2E12.jpg','jpg');
subplot(2,2,4)
E=A(2:2:row,2:2:col);
imshow(uint8(E));
imwrite(uint8(E),'exo2E22.jpg','jpg');

```

## 3. Quantification

```

clear all

A=imread('lena512.bmp');
colormap(gray(256));
A=double(A);
[row,col]=size(A)

```

```

for k=[4,5,6,7]
figure(k)
subplot(1,2,1)
imshow(uint8(A));
title ( "Original" );
subplot(1,2,2)
B=idivide(A,2.^k).*2.^k;

```

## 4. Détection des bords

```

clear all

A=imread('lena512.bmp');

```

```

colormap(gray(256));
A=double(A);
[row,col]=size(A)

```

```
% partial_x
G1(1,:)=(A(2,:)-A(1,:));
G1(2:row-1,:)=(A(3:row,:)-A(1:row-2,:))/2;
G1(row,:)=(A(row,:)-A(row-1,:));

% partial_y
G2(:,1)=(A(:,2)-A(:,1));
G2(:,2:col-1)=(A(:,3:col)-A(:,1:col-2))/2;
G2(:,col)=(A(:,col)-A(:,col-1));

% norme 2 du gradient
G=sqrt(G1.^2+G2.^2);

% negatif
G=255-G;

% normalisation sur [0;255]
m=min(min(G))
M=max(max(G))
Gn=255/(M-m).*(G-m);

subplot(1,3,1)
imshow(uint8(A));
title ("Original");
subplot(1,3,2)
imshow(uint8(G));
title ("Gradient (negatif)");
imwrite(uint8(G), 'exo41.jpg', 'jpg');
subplot(1,3,3)
imshow(uint8(Gn));
title ("Gradient normalise (negatif)");
imwrite(uint8(Gn), 'exo42.jpg', 'jpg');
```

## 5. Floutage

```
clear all

A=imread('lena512.bmp');
colormap(gray(256));
A=double(A);
[row,col]=size(A)

AA=A;
for t=1:100
    % partial_xx
    G1(1,:)=0*AA(1,:);
    G1(2:row-1,:)=AA(3:row,:)-2*AA(2:row-1,:)+AA(1:row-2,:);
    G1(row,:)=0*AA(row,:);
    % partial_yy
    G2(:,1)=0*AA(:,1);
    G2(:,2:col-1)=AA(:,3:col)-2*AA(:,2:col-1)+AA(:,1:col-2);
    G2(:,col)=0*AA(:,col);
    G=AA+(G1+G2)*0.2;
    % on se ramne a [0;255]
    m=min(min(G));
    M=max(max(G));
    G=255/(M-m).*(G-m);
    AA=G;
end

subplot(1,2,1)
imshow(uint8(A));
title ("Original");
subplot(1,2,2)
imshow(uint8(G));
title ("Floutage");
imwrite(uint8(G), 'exo5.jpg', 'jpg');
```

### ★ Exercice 1.24 (Traitement mathématique des images numériques)

Dans cette exercice nous allons nous intéresser à la manipulation d'images. Nous utiliserons des méthodes basées sur l'algèbre linéaire et l'analyse matricielle.

**Les pixels d'une image :** une image numérique en niveaux de gris (*grayscale* image en anglais) est un tableau de valeurs. Chaque case de ce tableau, qui stocke une valeur, se nomme un pixel. En notant  $n$  le nombre de lignes et  $p$  le nombre de colonnes de l'image, on manipule ainsi un tableau de  $n \times p$  pixels.

La figure ci-dessous montre une visualisation d'un tableau carré avec  $n = p = 512$ , ce qui représente  $512 \times 512 = 2^{18} = 262\,144$  pixels. Les appareils photos numériques peuvent enregistrer des images beaucoup plus grandes, avec plusieurs millions de pixels.

Les valeurs des pixels sont enregistrées dans l'ordinateur ou l'appareil photo numérique sous forme de nombres entiers entre 0 et 255, ce qui fait 256 valeurs possibles pour chaque pixel. La valeur 0 correspond au noir et la valeur 255 correspond au blanc. Les valeurs intermédiaires correspondent à des niveaux de gris allant du noir au blanc.

Pour transformer une image en une matrice il suffit d'indiquer dans notre script :

```
A=imread('flower.png');
colormap(gray(256));
A=double(A);
[row,col]=size(A)
```

```
colormap(gray(256));
imshow(uint8(A));
% uint8(x) convert x to unsigned 8-bit integer type
```

Octave la transforme en matrice avec la fonction `imread`. On a bien une matrice de taille  $512 \times 512$ . On peut voir une fleur avec :



FIGURE 1.7 – Flower (original)

### 1. Manipulations élémentaires

- 1.1. En utilisant une manipulation élémentaire de la matrice (sans faire de boucles et sans utiliser de fonctions) obtenir l'image 1.8a
- 1.2. On peut modifier le contraste en "mappant" par une fonction croissante plus rapidement ou plus lentement que la fonction identité  $i: [0;255] \rightarrow [0;255]$ ,  $i(g) = g$ . Par exemple, pour avoir une image plus claire, on pourra utiliser la fonction

$$\begin{aligned} f: [0;255] &\rightarrow [0;255] \\ g &\mapsto 255 + \frac{(g - 255)^3}{255^2} \end{aligned}$$

Appliquer cette transformation pour obtenir l'image 1.8b.

- 1.3. Pour obtenir l'image en négatif il suffit de prendre le complémentaire par rapport à 255

$$\begin{aligned} f: [0;255] &\rightarrow [0;255] \\ g &\mapsto 255 - g \end{aligned}$$

Appliquer cette transformation pour obtenir l'image 1.8c.

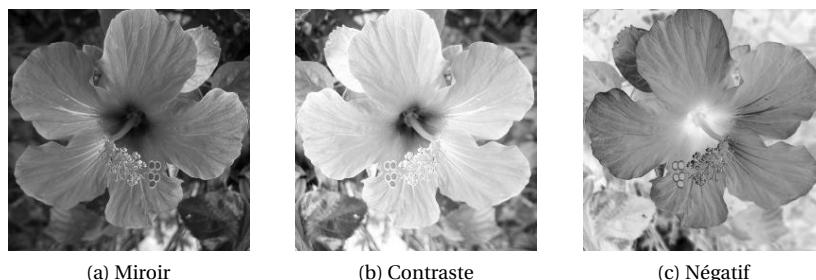


FIGURE 1.8 – Manipulations élémentaires

### 2. Résolution

Une matrice de taille  $2^9 \times 2^9$  contient  $2^{18}$  entiers ce qui prend pas mal de place en mémoire. On s'intéresse à des méthodes qui permettent d'être plus économique sans pour cela diminuer la qualité esthétique de l'image. Afin de réduire la place de stockage d'une image, on peut réduire sa résolution, c'est-à-dire diminuer le nombre de pixels. La façon la plus simple d'effectuer cette réduction consiste à supprimer des lignes et des colonnes dans l'image de départ. La figure suivante montre ce que l'on obtient si l'on retient que les lignes et les colonnes d'indices multiples de 4 ce qui donne une matrice  $2^6 \times 2^6$ . Appliquer cette transformation pour obtenir l'image suivante



### 3. Quantification

Une autre façon de réduire la place mémoire nécessaire pour le stockage consiste à utiliser moins de nombres entiers pour chaque valeur. On peut par exemple utiliser uniquement des nombres entiers entre 0 et 3, ce qui donnera une image avec uniquement 4 niveaux de gris. Une telle opération se nomme quantification.

On peut effectuer une conversion de l'image d'origine vers une image avec  $2^{8-k}$  niveaux de valeurs en effectuant les remplacements suivant : tous les valeurs entre 0 et  $2^k$  sont remplacées par la valeur 0, puis tous les valeurs entre  $2^k$  et  $2^{k+1}$  sont remplacées par la valeur  $2^k$  etc. Appliquer cette transformation pour obtenir les images suivantes :

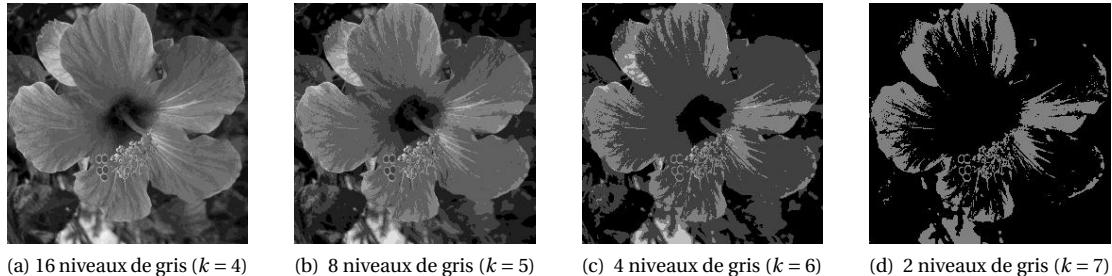


FIGURE 1.9 – Quantification

### 4. Détection des bords

Afin de localiser des objets dans les images, il est nécessaire de détecter les bords de ces objets. Ces bords correspondent à des zones de l'image où les valeurs des pixels changent rapidement. C'est le cas par exemple lorsque l'on passe du chapeau (qui est clair, donc avec des valeurs grandes) à l'arrière plan (qui est sombre, donc avec des valeurs petites). Afin de savoir si un pixel avec une valeur est le long d'un bord d'un objet, on prend en compte les valeurs de ses quatre voisins (deux horizontalement et deux verticalement). Pour cela nous allons calculer et afficher la norme d'un gradient discret en chaque pixel comme suit :

$$N(A_{i,j}) = \sqrt{(\partial_x A_{i,j})^2 + (\partial_y A_{i,j})^2}$$

avec

$$\text{pour } j = 1, \dots, 512, \quad \partial_x A_{i,j} \simeq \begin{cases} A_{i+1,j} - A_{i,j}, & \text{pour } i = 1 \\ \frac{A_{i+1,j} - A_{i-1,j}}{2}, & \text{pour } i = 2, \dots, 511 \\ A_{i,j} - A_{i-1,j}, & \text{pour } i = 512 \end{cases}$$

$$\text{pour } i = 1, \dots, 512, \quad \partial_y A_{i,j} \simeq \begin{cases} A_{i,j+1} - A_{i,j}, & \text{pour } j = 1 \\ \frac{A_{i,j+1} - A_{i,j-1}}{2}, & \text{pour } j = 2, \dots, 511 \\ A_{i,j} - A_{i,j-1}, & \text{pour } j = 512 \end{cases}$$

Appliquer cette transformation suivie de la transformation en négatif pour obtenir l'image 1.10a. On remarque que les valeurs obtenues appartiennent à l'intervalle  $[m = 133.73; M = 255]$ .

Pour améliorer le rendu, ramener le niveaux de gris à l'intervalle  $[0; 255]$  par une transformation affine ce qui donne l'image 1.10b. Cela correspond à la transformation

$$\begin{aligned} f: [m; M] &\rightarrow [0; 255] \\ g &\mapsto \frac{255}{M-m}(x-m) \end{aligned}$$

### 5. Floutage par diffusion

On veut lisser les endroits à fort gradient. Pour cela nous allons calculer une moyenne en chaque pixel comme suit :

$$\mathbb{A} \leftarrow \mathbb{A} + \frac{\partial_{xx}\mathbb{A} + \partial_{yy}\mathbb{A}}{5}$$

avec

$$\text{pour } j = 1, \dots, 512, \quad \partial_{xx} A_{i,j} \simeq \begin{cases} 0, & \text{pour } i = 1 \\ A_{i+1,j} - 2A_{i,j} + A_{i-1,j}, & \text{pour } i = 2, \dots, 511 \\ 0, & \text{pour } i = 512 \end{cases}$$

$$\text{pour } i = 1, \dots, 512, \quad \partial_{yy} A_{i,j} \simeq \begin{cases} 0, & \text{pour } j = 1 \\ A_{i,j+1} - 2A_{i,j} + A_{i,j-1}, & \text{pour } j = 2, \dots, 511 \\ 0, & \text{pour } j = 512 \end{cases}$$

Appliquer 100 fois cette transformation à la matrice  $\mathbb{A}$  pour obtenir l'image 1.10c.<sup>7</sup>

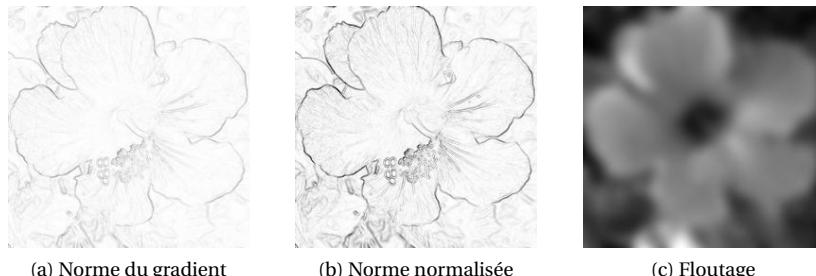


FIGURE 1.10 – Détection des bords et Floutage

## Correction

### 1. Manipulations élémentaires

#### 1.1. Miroir

```
clear all
A=imread('flower.png');
colormap(gray(256));
A=double(A);
[row,col]=size(A)
subplot(1,2,1)
```

```
imshow(uint8(A));
title ( "Original" );
subplot(1,2,2)
B=A(:,row:-1:1);
imshow(uint8(B));
title ( "Miroir" );
imwrite(uint8(B),'exo11.jpg','jpg');
```

#### 1.2. Contraste

```
clear all
A=imread('flower.png');
colormap(gray(256));
A=double(A);
[row,col]=size(A)
subplot(1,2,1)
```

```
imshow(uint8(A));
title ( "Original" );
subplot(1,2,2)
B=255+(A-255).^3/255^2;
imshow(uint8(B));
title ( "Clair" );
imwrite(uint8(B),'exo12.jpg','jpg');
```

#### 1.3. Négatif

```
clear all
A=imread('flower.png');
colormap(gray(256));
A=double(A);
[row,col]=size(A)
subplot(1,2,1)
```

```
imshow(uint8(A));
title ( "Original" );
subplot(1,2,2)
B=255-A;
imshow(uint8(B));
title ( "Négatif" );
imwrite(uint8(B),'exo13.jpg','jpg');
```

### 2. Résolution

```
clear all
A=imread('flower.png');
```

```
colormap(gray(256));
A=double(A);
[row,col]=size(A)
```

7. Cela correspond à un schéma 5 points explicite appliqué à l'équation de la chaleur  $\partial_t A = \nabla \cdot (f(\nabla A))$  avec  $f$  l'identité

```
E=A(1:4:row,1:4:col);
imshow(uint8(E));
imwrite(uint8(E), 'exo2E11.jpg', 'jpg');
```

### 3. Quantification

```
clear all
A=imread('flower.png');
colormap(gray(256));
A=double(A);
[row,col]=size(A)

for k=[4,5,6,7]
    figure(k)
    subplot(1,2,1)
    imshow(uint8(A));
    title ('Original');
    subplot(1,2,2)
    B=idivide(A,2^k)*2^k;
```

### 4. Détection des bords

```
clear all
% negatif
G=255-G;

% normalisation sur [0;255]
m=min(min(G))
M=max(max(G))
Gn=255/(M-m).*(G-m);

subplot(1,3,1)
imshow(uint8(A));
title ('Original');
subplot(1,3,2)
imshow(uint8(G));
title ('Gradient (negatif)');
imwrite(uint8(G), 'exo41.jpg', 'jpg');
subplot(1,3,3)
imshow(uint8(Gn));
title ('Gradient normalisé (negatif)');
imwrite(uint8(Gn), 'exo42.jpg', 'jpg');
```

### 5. Floutage

```
clear all
A=imread('flower.png');
colormap(gray(256));
A=double(A);
[row,col]=size(A)

AA=A;
for t=1:100
    % partial_xx
    G1(:,1)=0*AA(:,1,:);
    G1(2:row-1,:)=AA(3:row,:)-2*AA(2:row-1,:)+AA(1:row-2,:);
    G1(row,:)=0*AA(row,:);
    % partial_yy
    G2(:,1)=0*AA(:,1,:);
    G2(:,2:col-1)=AA(:,3:col)-2*AA(:,2:col-1)+AA(:,1:col-2);
    G2(:,col)=0*AA(:,col);
    G=AA+(G1+G2)*0.2;
    % on se ramne a [0;255]
    m=min(min(G));
    M=max(max(G));
    G=255/(M-m).*(G-m);
    AA=G;
end

subplot(1,2,1)
imshow(uint8(A));
title ('Original');
subplot(1,2,2)
imshow(uint8(G));
title ('Floutage');
imwrite(uint8(G), 'exo5.jpg', 'jpg');
```

## Calculs avec des nombres à virgule flottante

### ★ Exercice 1.25

Exécuter les instructions suivantes et commenter :

```
a=1;
b=1;
while a+b ~= a
    b=b/2
end
```

**Correction**

La variable  $b$  est divisée par deux à chaque étape tant que la somme de  $a$  et  $b$  demeure différente ( $\sim=$ ) de  $a$ . Si on opérait sur des nombres réels, ce programme ne s'arrêterait jamais, tandis qu'ici, il s'interrompt après un nombre fini d'itérations et renvoie la valeur suivante pour  $b$  :  $1.1102e-16 = \epsilon_M/2$ . Il existe donc au moins un nombre  $b$  différent de 0 tel que  $a + b = a$ .

**★ Exercice 1.26**

Écrire un programme pour calculer la suite

$$\begin{cases} I_1 = 1 - \frac{1}{e}, \\ I_{n+1} = 1 - (n+1)I_n, \text{ pour } n \in \mathbb{N}^*. \end{cases}$$

Comparer le résultat numérique avec la limite exacte  $I_n \rightarrow 0$  pour  $n \rightarrow +\infty$ .

**Correction**

La suite obtenue avec le programme ci-dessous ne tend pas vers zéro quand  $n$  tend vers l'infini : son signe alterne et elle diverge. Ce comportement est une conséquence directe de la propagation des erreurs d'arrondi.

```
I(1)=1-1/exp(1);
for n=[1:10]
    I(n+1)=1-(n+1)*I(n);
end
I(11)
```

**Fonctions trigonométrique****★ Exercice 1.27**

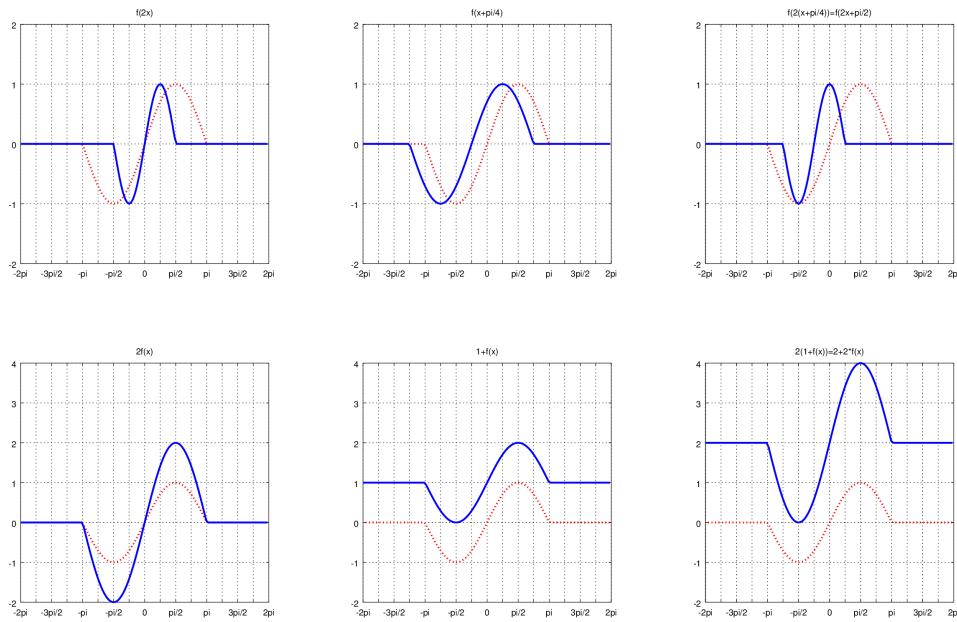
Soit  $f: \mathbb{R} \rightarrow \mathbb{R}$  définie par

$$f(x) = \begin{cases} \sin(x) & \text{si } x \in [-\pi; \pi], \\ 0 & \text{sinon.} \end{cases}$$

Tracer le graphe des fonctions  $f(2x)$ ,  $f\left(x + \frac{\pi}{4}\right)$ ,  $f\left(2\left(x + \frac{\pi}{4}\right)\right) = f\left(2x + \frac{\pi}{2}\right)$ ,  $2f(x)$ ,  $1 + f(x)$ ,  $2 + 2f(x) = 2(1 + f(x))$ .

**Correction**

On peut visualiser les différentes fonctions avec Octave comme suit :



```

x=[-2*pi:0.1:2*pi] ;

f=@(x) [ (x>-pi).* (x<pi) .* sin(x) ]
yf=f(x);

subplot(2,3,1)
yf1=f(2*x);
plot(x,yf,'r:','LineWidth',2,x,yf1,'b-','LineWidth',2)
axis([-2*pi 2*pi -2 2])
%xtick=[-2*pi:pi/4:2*pi];
%set(gca,'xtick',xtick);
set(gca,'XTick',-2*pi:pi/4:2*pi)
set(gca,'XTickLabel',{'-2pi','',''-3pi/2','',''-pi','',''-pi/2','','0','','pi/2','','pi','','3pi/2','','2pi'});
grid
title('f(2x)');

subplot(2,3,2)
yf2=f(x+pi/4);
plot(x,yf,'r:','LineWidth',2,x,yf2,'b-','LineWidth',2)
axis([-2*pi 2*pi -2 2])
set(gca,'XTick',-2*pi:pi/4:2*pi)
set(gca,'XTickLabel',{'-2pi','',''-3pi/2','',''-pi','',''-pi/2','','0','','pi/2','','pi','','3pi/2','','2pi'});
grid
title('f(x+pi/4)');

subplot(2,3,3)
yf3=f(2*x+pi/2);
plot(x,yf,'r:','LineWidth',2,x,yf3,'b-','LineWidth',2)
axis([-2*pi 2*pi -2 2])
set(gca,'XTick',-2*pi:pi/4:2*pi)
set(gca,'XTickLabel',{'-2pi','',''-3pi/2','',''-pi','',''-pi/2','','0','','pi/2','','pi','','3pi/2','','2pi'});
grid
title('f(2(x+pi/4))=f(2x+pi/2)');

subplot(2,3,4)
yf4=2*f(x);

```

```

plot(x,yf,'r:','LineWidth',2,x,yf4,'b-','LineWidth',2)
axis([-2*pi 2*pi -2 4])
set(gca,'XTick',-2*pi:pi/4:2*pi)
set(gca,'XTickLabel',{'-2pi','','-3pi/2','','-pi','','-pi/2','','0','','pi/2','','pi','','3pi/2','','2pi'})
grid
title('2f(x)');

subplot(2,3,5)
yf5=1+f(x);
plot(x,yf,'r:','LineWidth',2,x,yf5,'b-','LineWidth',2)
axis([-2*pi 2*pi -2 4])
set(gca,'XTick',-2*pi:pi/4:2*pi)
set(gca,'XTickLabel',{'-2pi','','-3pi/2','','-pi','','-pi/2','','0','','pi/2','','pi','','3pi/2','','2pi'})
grid
title('1+f(x)');

subplot(2,3,6)
yf6=2+2*f(x);
plot(x,yf,'r:','LineWidth',2,x,yf6,'b-','LineWidth',2)
axis([-2*pi 2*pi -2 4])
set(gca,'XTick',-2*pi:pi/4:2*pi)
set(gca,'XTickLabel',{'-2pi','','-3pi/2','','-pi','','-pi/2','','0','','pi/2','','pi','','3pi/2','','2pi'})
grid
title('2(1+f(x))=2+2*f(x)');

```

## Nombres complexes

### ★ Exercice 1.28

- ① Exécuter les instructions suivantes et commenter :

```
(-5)^(1/3)
```

- ② Montrer que  $i^i$  est un nombre réel, puis vérifier ce résultat avec Octave.

### Correction

- ① Octave sélectionne la première racine rencontrée en balayant le plan complexe dans le sens inverse des aiguilles d'une montre en partant de l'axe des réels :

```

>> (-5)^(1/3)
ans = 0.854987973338349 + 1.480882609682364i

```

Pour obtenir toutes les racines on peut lui demander de calculer toutes les racines du polynôme  $x^3 + 5$  :

```

pol = [1, 0, 0, 5]
roots(pol)
ans =
-1.70998 + 0.00000i
0.85499 + 1.48088i
0.85499 - 1.48088i

```

- ②  $i = e^{i\pi/2}$  et  $(f(x))^g(x) = e^{g(x)\ln(f(x))}$  donc  $i^i = e^{i\ln(i)} = e^{i\ln(e^{i\pi/2})} = e^{i^2\pi/2} = e^{-\pi/2} \in \mathbb{R}$  :

```

>> i^i
ans = 0.207879576350762
>> exp(-pi/2)
ans = 0.207879576350762

```

### Exercice 1.29

Calculer les racines carrées de

- a) 1                    b)  $i$                     c)  $3 + 4i$                     d)  $8 - 6i$                     e)  $7 + 24i$                     f)  $3 - 4i$                     g)  $24 - 10i$   
 h)  $\frac{1+i}{\sqrt{2}}$                     i) 2                    j)  $3i$                     k)  $2+3i$                     l)  $31+bi$  ( $b$  est un paramètre réel)

**Correction**

- (a)  $z = 1 + 0i \implies x = 1, y = 0 \implies \begin{cases} \alpha = 1 \\ \beta = 0 \end{cases} \implies w_0 = 1 \text{ et } w_1 = -1$
- (b)  $z = i \implies x = 0, y = 1 \implies \begin{cases} \alpha = \frac{1}{\sqrt{2}} \\ \beta = \frac{1}{\sqrt{2}} \end{cases} \implies w_0 = \frac{1}{\sqrt{2}} + i \frac{1}{\sqrt{2}} = e^{i\pi/4} \text{ et } w_1 = -\frac{1}{\sqrt{2}} - i \frac{1}{\sqrt{2}} = e^{i5\pi/4}$
- (c)  $z = 3 + 4i \implies x = 3, y = 4 \implies \begin{cases} \alpha = 2 \\ \beta = 1 \end{cases} \implies w_0 = 2 + i \text{ et } w_1 = -2 - i$
- (d)  $z = 8 - 6i \implies x = 8, y = -6 \implies \begin{cases} \alpha = 3 \\ \beta = -1 \end{cases} \implies w_0 = 3 - i \text{ et } w_1 = -3 + i : \text{on aurait pu tout simplement remarquer que } z \text{ est le conjugué du complexe au point (c)}$
- (e)  $z = 7 + 24i \implies x = 7, y = 24 \implies \begin{cases} \alpha = 4 \\ \beta = 3 \end{cases} \implies w_0 = 4 + 3i \text{ et } w_1 = -4 - 3i$
- (f)  $z = 3 - 4i \implies x = 3, y = -4 \implies \begin{cases} \alpha = 2 \\ \beta = -1 \end{cases} \implies w_0 = 2 - i \text{ et } w_1 = -2 + i$
- (g)  $z = 24 - 10i \implies x = 24, y = -10 \implies \begin{cases} \alpha = 5 \\ \beta = -1 \end{cases} \implies w_0 = 5 - i \text{ et } w_1 = -5 + i$
- (h)  $z = \frac{1+i}{\sqrt{2}} = \frac{1}{2}e^{i\pi/4} \implies w_0 = \frac{1}{\sqrt{2}}e^{i\pi/8} \text{ et } w_1 = \frac{1}{\sqrt{2}}e^{i(\pi/8+\pi)}$
- (i)  $z = 2 = 2e^{i0} \implies w_0 = \sqrt{2}e^{i0} = \sqrt{2} \text{ et } w_1 = \sqrt{2}e^{i\pi} = -\sqrt{2}$
- (j)  $z = 3i = 3e^{i\pi} \implies w_0 = \sqrt{3}e^{i\pi/4} \text{ et } w_1 = \sqrt{3}e^{i5\pi/4}$
- (k)  $z = 2 + 3i \implies x = 2, y = 3 \implies \begin{cases} \alpha = \sqrt{\frac{2+\sqrt{13}}{2}} \\ \beta = \sqrt{\frac{-2+\sqrt{13}}{2}} \end{cases} \implies w_0 = \sqrt{\frac{2+\sqrt{13}}{2}} + i\sqrt{\frac{-2+\sqrt{13}}{2}} \text{ et } w_1 = -w_0$
- (l)  $z = 1 + bi$  ( $b$  est un paramètre réel)  $\implies x = 1, y = b \implies \begin{cases} \alpha = \sqrt{\frac{1+\sqrt{1+b^2}}{2}} \\ \beta = \operatorname{sgn}(b)\sqrt{\frac{-1+\sqrt{1+b^2}}{2}} \end{cases} \implies$
- $$\begin{cases} w_0 = \sqrt{\frac{1+\sqrt{1+b^2}}{2}} + i\sqrt{\frac{-1+\sqrt{1+b^2}}{2}} & \text{si } b > 0, \\ w_0 = \sqrt{\frac{1+\sqrt{1+b^2}}{2}} - i\sqrt{\frac{-1+\sqrt{1+b^2}}{2}} & \text{si } b < 0, \\ w_0 = 1 & \text{si } b = 0, \end{cases}$$
- et  $w_1 = -w_0$

```
roots([1, 0, -(1)])
roots([1, 0, -(i)])
roots([1, 0, -(3+4*i)])
roots([1, 0, -(8-6*i)])
```

```
roots([1, 0, -(7+24*i)])
roots([1, 0, -(3-4*i)])
roots([1, 0, -(24-10*i)])
roots([1, 0, -(1+i)/sqrt(2))])
```

```
roots([1, 0, -(2)])
roots([1, 0, -(3*i)])
roots([1, 0, -(2+3*i)])
```

**Exercice 1.30**

Trouver (sous forme algébrique et trigonométrique) toutes les racines complexes de l'équation

$$z^2 - (1 + 2i)z + i - 1 = 0.$$

**Correction**

$A = 1, b = -1 - 2i, c = i - 1, \Delta = b^2 - 4ac = (-(1 + 2i))^2 - 4(i - 1) = 1$  donc  $\delta_0 = 1$  et  $\delta_1 = -\delta_0$  et finalement  $z_0 = \frac{-b + \delta_0}{2} = 1 + i = \sqrt{2} \left( \cos\left(\frac{\pi}{4}\right) + i \sin\left(\frac{\pi}{4}\right) \right)$  et  $z_1 = \frac{-b - \delta_0}{2} = i = \cos\left(\frac{\pi}{2}\right) + i \sin\left(\frac{\pi}{2}\right)$ .

```
roots([1, -(1+2*i), i-1])
```

### Exercice 1.31

Trouver toutes les solutions complexes des équations

- |                                   |                                      |                                    |
|-----------------------------------|--------------------------------------|------------------------------------|
| a) $z^2 - 2iz + 2 - 4i = 0$ ,     | b) $z^2 + 2z + (1 - 2i) = 0$ ,       | c) $z^2 + 2(1+i)z - 5(1+2i) = 0$ , |
| d) $z^2 - (5+3i)z + 7i + 4 = 0$ , | e) $z^2 - \sqrt{3}z - i = 0$ ,       | f) $z^2 + z + 1 = 0$ ,             |
| g) $z^2 - (1+2i)z + i - 1 = 0$ ,  | h) $z^2 - (5-14i)z - 2(5i+12) = 0$ , | i) $z^2 - (3+4i)z - 1 + 5i = 0$ ,  |
| j) $4z^2 - 2z + 1 = 0$ ,          | k) $z^2 - (11-5i)z + 24 - 27i = 0$ , | l) $z^4 + 2z^2 + 4 = 0$ .          |

#### Correction

1.  $z^2 - 2iz + 2 - 4i = 0 \implies \delta^2 = (-2i)^2 - 4(2-4i) = -12 + 16i \implies \delta_0 = 2+4i \implies z_0 = \frac{2i+2+4i}{2} = 1+3i \text{ et } z_1 = \frac{2i-2-4i}{2} = -1-i$
2.  $z^2 + 2z + (1 - 2i) = 0 \implies \delta^2 = 2^2 - 4(1 - 2i) = 8i \implies \delta_0 = 2+2i \implies z_0 = \frac{-2+2+2i}{2} = i \text{ et } z_1 = \frac{-2-2-2i}{2} = -2-i$
3.  $z^2 + 2(1+i)z - 5(1+2i) = 0 \implies \delta^2 = (2(1+i))^2 - 4(-5(1+2i)) = 20+48i \implies \delta_0 = 6+4i \implies z_0 = \frac{-2-2i+6+4i}{2} = 2+i \text{ et } z_1 = \frac{-2-2i-6-4i}{2} = -4-3i$
4.  $z^2 - (5+3i)z + 7i + 4 = 0 \implies \delta^2 = (-(5+3i))^2 - 4(4+7i) = 2i \implies \delta_0 = 1+i \implies z_0 = \frac{5+3i+1+i}{2} = 3+2i \text{ et } z_1 = \frac{5+3i-1-i}{2} = 2+i$
5.  $z^2 - \sqrt{3}z - i = 0 \implies \delta^2 = (-\sqrt{3})^2 - 4(-i) = 3+4i \implies \delta_0 = 2+i \implies z_0 = \frac{\sqrt{3}+2+i}{2} \text{ et } z_1 = \frac{\sqrt{3}-2-i}{2}$
6.  $z^2 + z + 1 = 0 \implies \delta^2 = -3 \implies \delta_0 = i\sqrt{3} \implies z_0 = \frac{-1+i\sqrt{3}}{2} \text{ et } z_1 = \frac{-1-i\sqrt{3}}{2}$
7.  $z^2 - (1+2i)z + i - 1 = 0 \implies \delta^2 = (-(1+2i))^2 - 4(i-1) = 1 \implies \delta_0 = 1 \implies z_0 = \frac{1+2i-1}{2} = i \text{ et } z_1 = \frac{1+2i+1}{2} = 1+i$
8.  $z^2 - (5-14i)z - 2(5i+12) = 0 \implies \delta^2 = -75 - 100i \implies \delta_0 = 5 - 10i \implies z_0 = 5 - 12i \text{ et } z_1 = -2i$
9.  $z^2 - (3+4i)z - 1 + 5i = 0 \implies \delta^2 = -3+4i \implies \delta_0 = 1+2i \implies z_0 = 2+3i \text{ et } z_1 = 1+i$
10.  $4z^2 - 2z + 1 = 0 \implies \delta^2 = -12 \implies \delta_0 = 2\sqrt{3}i \implies z_0 = \frac{1-\sqrt{3}i}{4} \text{ et } z_1 = \frac{1+\sqrt{3}i}{4}$
11.  $z^2 - (11-5i)z + 24 - 27i = 0 \implies \delta^2 = -2i \implies \delta_0 = 1-i \implies z_0 = 6-3i \text{ et } z_1 = 5-2i$
12. On pose  $w = z^2$  et on résout d'abord  $w^2 + 2w + 4 = 0 \implies \delta^2 = -12 \implies \delta_0 = 2\sqrt{3}i \implies w_0 = -1 + \sqrt{3}i \text{ et } w_1 = -1 - \sqrt{3}i$ .  
On résout maintenant les deux équations

$$\star \ z^2 = w_0 \implies \begin{cases} \alpha = \frac{1}{\sqrt{2}} \\ \beta = \frac{\sqrt{3}}{\sqrt{2}} \end{cases} \implies z_0 = \frac{1}{\sqrt{2}} + i\frac{\sqrt{3}}{\sqrt{2}} \text{ et } z_1 = -\frac{1}{\sqrt{2}} - i\frac{\sqrt{3}}{\sqrt{2}}$$

$$\star \ z^2 = w_1 \implies \begin{cases} \alpha = \frac{1}{\sqrt{2}} \\ \beta = -\frac{\sqrt{3}}{\sqrt{2}} \end{cases} \implies z_3 = \frac{1}{\sqrt{2}} - i\frac{\sqrt{3}}{\sqrt{2}} \text{ et } z_4 = -\frac{1}{\sqrt{2}} + i\frac{\sqrt{3}}{\sqrt{2}}$$

```
roots([1,-2*i,2-4*i])
roots([1,2,1-2*i])
roots([1,2*(1+i),-5*(1+2*i)])
roots([1,-(5+3*i),7*i+4])
```

```
roots([1,-sqrt(3),-i])
roots([1,1,1])
roots([1,-(1+2*i),i-1])
roots([1,-(5-14*i),-2*(5*i+12)])
```

```
roots([1,-(3+4*i),-1+5*i])
roots([4,-2,1])
roots([1,-(11-5*i),24-27*i])
roots([4,0,2,0,4])
```

## Racines de polynômes et Division euclidienne

### Exercice 1.32

Soit  $p$  un polynôme de degré 4. On sait qu'il assume en  $x = 2$  et  $x = 3$  sa valeur maximale 3 et qu'il s'annule en  $x = 1$ . Calculer  $p(4)$ .

#### Correction

- \*  $p(2) = p(3) = 3$  et  $p$  a degré 4 donc  $p(x) - 3 = (x-2)(x-3)q(x)$  avec  $q$  un polynôme de degré 2.
- \*  $x = 2$  et  $x = 3$  sont des maxima pour  $p$ , donc  $p'(2) = p'(3) = 0$ . Comme  $p'(x) = (2x-5)q(x) + (x-2)(x-3)q'(x)$ , alors  $q(2) = 0$  et  $q(3) = 0$ , ce qui implique  $q(x) = a(x-2)(x-3)$ . On obtient alors  $p(x) = 3 + a(x-2)^2(x-3)^2$ .
- \* Comme  $p(1) = 0$  alors  $a = -3/4$  et on conclut que  $p(4) = 0$ .

### Exercice 1.33

Soit  $n \in \mathbb{N}^*$ . Démontrer que  $P = x^{n+1} + x^n - 2x^{n-1} + nx - n$  est divisible par  $(x-1)$ .

**Correction**

Un polynôme  $P$  est divisible par  $(x - 1)$  si  $x = 1$  est racine de  $P$ , donc il suffit de vérifier que  $P(1) = 0$ .

**Exercice 1.34**

Soit  $n \in \mathbb{N}^*$ . Démontrer que  $P = (x - 2)^{2n} + (x - 1)^n - 1$  est divisible par  $x^2 - 3x + 2$ .

**Correction**

Comme  $x^2 - 3x + 2 = (x - 1)(x - 2)$ , il suffit de montrer que  $x = 1$  et  $x = 2$  sont racines du polynôme  $P$ , autrement dit que  $P(1) = P(2) = 0$ .

**Exercice 1.35**

Soit  $n \in \mathbb{N}$ ,  $n \geq 2$ . On considère le polynôme  $P$  défini par  $P = ax^{n+1} + bx^n + 1$ . Déterminer les réels  $a$  et  $b$  pour que  $P$  soit divisible par  $(x - 1)^2$ .

**Correction**

Si  $P$  est divisible par  $(x - 1)^2$ , cela signifie que  $x = 1$  est racine au moins double, donc il faut que  $P(1) = P'(1) = 0$ . Comme  $P(1) = a + b + 1$  et  $P'(1) = a(n + 1) + bn$ , on conclut que  $a = n$  et  $b = -n - 1$ .

**Exercice 1.36 (multiplicité)**

Calculer la multiplicité de la racine  $x_0$  de  $P$  dans les cas suivants :

- \*  $x_0 = 1$  et  $P(x) = x^4 - x^3 - 3x^2 + 5x - 2$ ,
- \*  $x_0 = i$  et  $P(x) = x^3 - ix^2 + x - i$ .

**Correction**

- \*  $x_0 = 1$  a multiplicité 3 car

$$\begin{array}{ll} P(x) = x^4 - x^3 - 3x^2 + 5x - 2 & P(1) = 0 \\ P'(x) = 4x^3 - 3x^2 - 6x + 5 & P'(1) = 0 \\ P''(x) = 12x^2 - 6x - 6 & P''(1) = 0 \\ P'''(x) = 24x - 6 & P'''(1) = 18 \end{array}$$

et  $P(x) = (x - 1)^3 Q(x)$  avec  $Q$  polynôme de degré 1 non divisible par  $(x - 1)$ .

```
p=[1 -1 -3 5 -2]
x0=1
valuation=polyval(p,x0)
n=0
while valuation==0
    p=polyder(p)
    valuation=polyval(p,x0)
    n+=1
end
```

- \*  $x_0 = i$  a multiplicité 2 car

$$\begin{array}{ll} P(x) = x^3 - ix^2 + x - i & P(1) = 0 \\ P'(x) = 3x^2 - 2ix + 1 & P'(1) = 0 \\ P''(x) = 6x - 2i & P''(1) = 4i \end{array}$$

et  $P(x) = (x - i)^2 Q(x)$  avec  $Q$  polynôme de degré 1 non divisible par  $(x - i)$ .

```
p=[1 -i 1 -i]
x0=i
valuation=polyval(p,x0)
n=0
while valuation==0
    p=polyder(p)
    valuation=polyval(p,x0)
    n+=1
end
```

### Exercice 1.37 (Division euclidienne)

Effectuer la division euclidienne de  $f$  par  $g$  dans les cas suivants

a)  $f(x) = 7x^4 - x^3 + 2x - 4$  et  $g(x) = x^2 - 3x + 5$ ,

b)  $f(x) = x^5 - x^4 - x + 1$  et  $g(x) = x^2 - 2x + 1$ .

#### Correction

$$f=[7 \ -1 \ 0 \ 2 \ -4]$$

$$g=[1 \ -3 \ 5]$$

$$[q,r]=\text{deconv}(f,g)$$

$$f=[1 \ -1 \ 0 \ 0 \ -1 \ 1]$$

$$g=[1 \ -2 \ 1]$$

$$[q,r]=\text{deconv}(f,g)$$

$$\begin{array}{r} 7x^4 \quad -x^3 \quad +2x \quad -4 \\ -7x^4+21x^3-35x^2 \\ \hline 20x^3-35x^2 \quad +2x \quad -4 \\ -20x^3+60x^2-100x \\ \hline 25x^2 \quad -98x \quad -4 \\ -25x^2 \quad +75x-125 \\ \hline -23x-129 \end{array}$$

$$\begin{array}{r} x^5 \quad -x^4 \quad & -x+1 \\ -x^5+2x^4 \quad -x^3 & \\ \hline x^4 \quad -x^3 \quad -x+1 \\ -x^4+2x^3 \quad -x^2 & \\ \hline x^3 \quad -x^2 \quad -x+1 \\ -x^3+2x^2 \quad -x & \\ \hline x^2-2x+1 \\ -x^2+2x-1 \\ \hline 0 \end{array}$$

donc  $f(x) = (7x^2 + 20x + 25)g(x) + (-23x - 129)$ .

donc  $f(x) = (x^3 + x^2 + x + 1)g(x)$ .

### Exercice 1.38

On considère le polynôme

$$P(x) := x^5 - 2x^4 + x^3 + x^2 - 2x + 1.$$

1. Écrire la formule de Taylor de  $P$  en 1. Montrer que 1 est une racine de  $P$  et en déterminer son ordre de multiplicité.
2. Factoriser  $P$  d'abord sur  $\mathbb{R}$ , puis sur  $\mathbb{C}$ .

#### Correction

1. Pour un polynôme  $P$  de degré 5, la formule de Taylor en 1 s'écrit

$$P(x) = \sum_{k=0}^{k=5} P^{(k)}(1) \frac{(x-1)^k}{k!}.$$

On calcul alors les dérivées  $P^{(k)}(x)$  pour  $k = 0, \dots, 5$  et on les évalue en 1 :

$$\begin{array}{ll} P(x) = x^5 - 2x^4 + x^3 + x^2 - 2x + 1 & P(1) = 0 \\ P'(x) = 5x^4 - 8x^3 + 3x^2 + 2x - 2 & P'(1) = 0 \\ P''(x) = 20x^3 - 24x^2 + 6x + 2 & P''(1) = 4 \\ P'''(x) = 60x^2 - 48x + 6 & P'''(1) = 18 \\ P^{IV}(x) = 120x - 48 & P^{IV}(1) = 72 \\ P^V(x) = 120 & P^V(1) = 120 \end{array}$$

La formule de Taylor en 1 s'écrit alors

$$\begin{aligned} P(x) &= 4 \frac{(x-1)^2}{2} + 18 \frac{(x-1)^3}{6} + 72 \frac{(x-1)^4}{24} + 120 \frac{(x-1)^5}{120} \\ &= 2(x-1)^2 + 3(x-1)^3 + 3(x-1)^4 + (x-1)^5 \\ &= (x-1)^2(x^3 + 1) \end{aligned}$$

Ceci implique que 1 est une racine de  $P$  de multiplicité 2.

2. On pose  $Q(x) = x^3 + 1$ . Une racine réelle évidente de  $Q$  est  $-1$ . En effectuant la division euclidienne de  $Q$  par  $(x+1)$  on obtient

$$Q(x) = (x+1)(x^2 - x + 1).$$

Étant donné que  $x^2 - x + 1$  ne se factorise pas sur  $\mathbb{R}$ , on conclut que la factorisation de  $P$  sur  $\mathbb{R}$  est

$$P(x) = (x-1)^2(x+1)(x^2 - x + 1).$$

Pour factoriser  $P$  sur  $\mathbb{C}$  il ne reste que factoriser le polynôme  $x^2 - x + 1$  sur  $\mathbb{C}$ . Avec les notations usuelles du cours, on a  $\Delta = -3$  et une racine carrée de  $\Delta$  est  $\delta = i\sqrt{3}$  d'où les deux racines complexes conjuguées  $x_1 = \frac{1}{2} - i\frac{\sqrt{3}}{2} = e^{i\frac{\pi}{3}}$  et  $x_2 = \frac{1}{2} + i\frac{\sqrt{3}}{2} = e^{i\frac{5\pi}{3}}$ . On conclut que la factorisation de  $P$  sur  $\mathbb{C}$  est

$$P(x) = (x-1)^2(x+1)(x - e^{i\frac{\pi}{3}})(x - e^{i\frac{5\pi}{3}}).$$

### Exercice 1.39

On considère le polynôme

$$P(x) := x^4 - x^3 - 3x^2 + 5x - 2.$$

1. Montrer que  $x = 1$  est une racine du polynôme et en déterminer son ordre de multiplicité.
2. Factoriser  $P$  à l'aide de la formule de Taylor.

#### Correction

1. On a

$$\begin{array}{ll} P(x) = x^4 - x^3 - 3x^2 + 5x - 2 & P(1) = 0 \\ P'(x) = 4x^3 - 3x^2 - 6x + 5 & P'(1) = 0 \\ P''(x) = 12x^2 - 6x - 6 & P(1) = 0 \\ P'''(x) = 24x - 6 & P'''(1) = 18 \\ P^{IV}(x) = 24 & P^{IV}(1) = 24 \end{array}$$

On en déduit que  $x = 1$  est une racine de multiplicité 3.

2. Pour un polynôme  $P$  de degré 4, la formule de Taylor en 1 s'écrit

$$P(x) = \sum_{k=0}^{k=4} P^{(k)}(1) \frac{(x-1)^k}{k!}.$$

La formule de Taylor en 1 s'écrit alors

$$P(x) = 18 \frac{(x-1)^3}{3!} + 24 \frac{(x-1)^4}{4!} = 3(x-1)^3 + (x-1)^4 = (x-1)^3(x+2).$$

### Exercice 1.40

On considère le polynôme

$$P(x) = x^5 - x^4 - 2x^3 + 5x^2 - 5x + 2.$$

1. Écrire la formule de Taylor de  $P$  en 1 ; montrer que 1 est une racine de  $P$  et déterminer son ordre de multiplicité.
2. Factoriser  $P$  d'abord sur  $\mathbb{R}$ , puis sur  $\mathbb{C}$  en sachant qu'il possède une autre racine réelle évidente (*i.e.* elle est un diviseur du terme de degré 0).

#### Correction

1. Pour un polynôme  $P$  de degré 5, la formule de Taylor en 1 s'écrit

$$P(x) = \sum_{k=0}^{k=5} P^{(k)}(1) \frac{(x-1)^k}{k!}.$$

On calcul alors les dérivées  $P^{(k)}(x)$  pour  $k = 0, \dots, 5$  et on les évalue en 1 :

$$P(x) = x^5 - x^4 - 2x^3 + 5x^2 - 5x + 2 \quad P(1) = 0$$

$$\begin{aligned}
 P'(x) &= 5x^4 - 4x^3 - 6x^2 + 10x - 5 & P'(1) &= 0 \\
 P''(x) &= 20x^3 - 12x^2 - 12x + 10 & P''(1) &= 6 \\
 P'''(x) &= 60x^2 - 24x - 12 & P'''(1) &= 24 \\
 P^{IV}(x) &= 120x - 24 & P^{IV}(1) &= 96 \\
 P^V(x) &= 120 & P^V(1) &= 120
 \end{aligned}$$

La formule de Taylor en 1 s'écrit alors

$$\begin{aligned}
 P(x) &= 6\frac{(x-1)^2}{2} + 24\frac{(x-1)^3}{6} + 96\frac{(x-1)^4}{24} + 120\frac{(x-1)^5}{120} \\
 &= 3(x-1)^2 + 4(x-1)^3 + 4(x-1)^4 + (x-1)^5 \\
 &= (x-1)^2(x^3 + x^2 - x + 2)
 \end{aligned}$$

Ceci implique que 1 est une racine de  $P$  de multiplicité 2.

2. On pose  $Q(x) = x^3 + x^2 - x + 2$ . Une racine réelle évidente de  $Q$  est  $-2$ . En effectuant la division euclidienne de  $Q$  par  $(x+2)$  on obtient

$$Q(x) = (x+2)(x^2 - x + 1).$$

Étant donné que  $x^2 - x + 1$  ne se factorise pas sur  $\mathbb{R}$ , on conclut que la factorisation de  $P$  sur  $\mathbb{R}$  est

$$P(x) = (x-1)^2(x+2)(x^2 - x + 1).$$

Pour factoriser  $P$  sur  $\mathbb{C}$  il ne reste que factoriser le polynôme  $x^2 - x + 1$  sur  $\mathbb{C}$ . Avec les notations usuelles du cours, on a  $\Delta = -3$  et une racine carrée de  $\Delta$  est  $\delta = i\sqrt{3}$  d'où les deux racines complexes conjuguées  $x_1 = \frac{1}{2} - i\frac{\sqrt{3}}{2} = e^{i\frac{\pi}{3}}$  et  $x_2 = \frac{1}{2} + i\frac{\sqrt{3}}{2} = e^{i\frac{5\pi}{3}}$ . On conclut que la factorisation de  $P$  sur  $\mathbb{C}$  est

$$P(x) = (x-1)^2(x+2)(x - e^{i\frac{\pi}{3}})(x - e^{i\frac{5\pi}{3}}).$$

### Exercice 1.41

On considère le polynôme

$$P(x) = x^5 - x^4 - 2x^3 + x^2 - x - 2.$$

1. Écrire la formule de Taylor de  $P$  en  $-1$ ; montrer que  $-1$  est une racine de  $P$  et déterminer son ordre de multiplicité.
2. Factoriser  $P$  d'abord sur  $\mathbb{R}$ , puis sur  $\mathbb{C}$  en sachant qu'il possède une autre racine réelle évidente (*i.e.* elle est un diviseur du terme de degré 0).

### Correction

1. Pour un polynôme  $P$  de degré 5, la formule de Taylor en  $-1$  s'écrit

$$P(x) = \sum_{k=0}^{k=5} P^{(k)}(-1) \frac{(x+1)^k}{k!}.$$

On calcul alors les dérivées  $P^{(k)}(x)$  pour  $k = 0, \dots, 5$  et on les évalue en  $-1$ :

$$\begin{aligned}
 P(x) &= x^5 - x^4 - 2x^3 + x^2 - x - 2 & P(-1) &= 0 \\
 P'(x) &= 5x^4 - 4x^3 - 6x^2 + 2x - 1 & P'(-1) &= 0 \\
 P''(x) &= 20x^3 - 12x^2 - 12x + 2 & P''(-1) &= -18 \\
 P'''(x) &= 60x^2 - 24x - 12 & P'''(-1) &= 72 \\
 P^{IV}(x) &= 120x - 24 & P^{IV}(-1) &= -144 \\
 P^V(x) &= 120 & P^V(-1) &= 120
 \end{aligned}$$

La formule de Taylor en  $-1$  s'écrit alors

$$\begin{aligned}
 P(x) &= -18\frac{(x+1)^2}{2} + 72\frac{(x+1)^3}{6} - 144\frac{(x+1)^4}{24} + 120\frac{(x+1)^5}{120} \\
 &= -9(x+1)^2 + 12(x+1)^3 - 6(x+1)^4 + (x+1)^5
 \end{aligned}$$

$$= (x+1)^2(x^3 - 3x^2 + 3x - 2)$$

Ceci implique que  $-1$  est une racine de  $P$  de multiplicité 2.

2. On pose  $Q(x) = x^3 - 3x^2 + 3x - 2$ . Une racine réelle évidente de  $Q$  est 2. En effectuant la division euclidienne de  $Q$  par  $(x-2)$  on obtient

$$Q(x) = (x-2)(x^2 - x + 1).$$

Étant donné que  $x^2 - x + 1$  ne se factorise pas sur  $\mathbb{R}$ , on conclut que la factorisation de  $P$  sur  $\mathbb{R}$  est

$$P(x) = (x+1)^2(x-2)(x^2 - x + 1).$$

Pour factoriser  $P$  sur  $\mathbb{C}$  il ne reste que factoriser le polynôme  $x^2 - x + 1$  sur  $\mathbb{C}$ . Avec les notations usuelles du cours, on a  $\Delta = -3$  et une racine carrée de  $\Delta$  est  $\delta = i\sqrt{3}$  d'où les deux racines complexes conjuguées  $x_1 = \frac{1}{2} - i\frac{\sqrt{3}}{2} = e^{i\frac{\pi}{3}}$  et  $x_2 = \frac{1}{2} + i\frac{\sqrt{3}}{2} = e^{i\frac{5\pi}{3}}$ . On conclut que la factorisation de  $P$  sur  $\mathbb{C}$  est

$$P(x) = (x+1)^2(x-2)(x - e^{i\frac{\pi}{3}})(x - e^{i\frac{5\pi}{3}}).$$

### Exercice 1.42

On considère le polynôme

$$P(x) = x^5 - 5x^4 + 10x^3 - 11x^2 + 7x - 2.$$

1. Écrire la formule de Taylor de  $P$  en 1 ; montrer que 1 est une racine de  $P$  et déterminer son ordre de multiplicité.
2. Factoriser  $P$  d'abord sur  $\mathbb{R}$ , puis sur  $\mathbb{C}$  en sachant qu'il possède une autre racine réelle évidente (*i.e.* elle est un diviseur du terme de degré 0).

#### Correction

1. Pour un polynôme  $P$  de degré 5, la formule de Taylor en 1 s'écrit

$$P(x) = \sum_{k=0}^{k=5} P^{(k)}(1) \frac{(x-1)^k}{k!}.$$

On calcul alors les dérivées  $P^{(k)}(x)$  pour  $k = 0, \dots, 5$  et on les évalue en 1 :

$$\begin{aligned} P(x) &= x^5 - 5x^4 + 10x^3 - 11x^2 + 7x - 2 & P(1) &= 0 \\ P'(x) &= 5x^4 - 20x^3 + 30x^2 - 22x + 7 & P'(1) &= 0 \\ P''(x) &= 20x^3 - 60x^2 + 60x - 22 & P''(1) &= -2 \\ P'''(x) &= 60x^2 - 120x + 60 & P'''(1) &= 0 \\ P^{IV}(x) &= 120x - 120 & P^{IV}(1) &= 0 \\ P^V(x) &= 120 & P^V(1) &= 120 \end{aligned}$$

La formule de Taylor en 1 s'écrit alors

$$\begin{aligned} P(x) &= -2 \frac{(x-1)^2}{2} + 120 \frac{(x-1)^5}{120} \\ &= (x-1)^2(x^3 - 3x^2 + 3x - 2) \end{aligned}$$

Ceci implique que 1 est une racine de  $P$  de multiplicité 2.

2. On pose  $Q(x) = x^3 - 3x^2 + 3x - 2$ . Une racine réelle évidente de  $Q$  est 2. En effectuant la division euclidienne de  $Q$  par  $(x-2)$  on obtient

$$Q(x) = (x-2)(x^2 - x + 1).$$

Étant donné que  $x^2 - x + 1$  ne se factorise pas sur  $\mathbb{R}$ , on conclut que la factorisation de  $P$  sur  $\mathbb{R}$  est

$$P(x) = (x-1)^2(x-2)(x^2 - x + 1).$$

Pour factoriser  $P$  sur  $\mathbb{C}$  il ne reste que factoriser le polynôme  $x^2 - x + 1$  sur  $\mathbb{C}$ . Avec les notations usuelles du cours, on a  $\Delta = -3$  et une racine carrée de  $\Delta$  est  $\delta = i\sqrt{3}$  d'où les deux racines complexes conjuguées  $x_1 = \frac{1}{2} - i\frac{\sqrt{3}}{2} = e^{i\frac{\pi}{3}}$  et

$x_2 = \frac{1}{2} + i\frac{\sqrt{3}}{2} = e^{i\frac{5\pi}{3}}$ . On conclut que la factorisation de  $P$  sur  $\mathbb{C}$  est

$$P(x) = (x-1)^2(x-2)(x-e^{i\frac{\pi}{3}})(x-e^{i\frac{5\pi}{3}}).$$

### Exercice 1.43

On considère le polynôme

$$P(x) = x^6 - 5x^5 + 10x^4 - 12x^3 + 11x^2 - 7x + 2.$$

1. Écrire la formule de Taylor de  $P$  en 1 ; montrer que 1 est une racine de  $P$  et déterminer son ordre de multiplicité.
2. Factoriser  $P$  d'abord sur  $\mathbb{R}$ , puis sur  $\mathbb{C}$  en sachant qu'il possède une autre racine réelle évidente (*i.e.* elle est un diviseur du terme de degré 0).

#### Correction

1. Pour un polynôme  $P$  de degré 6, la formule de Taylor en 1 s'écrit

$$P(x) = \sum_{k=0}^{k=6} P^{(k)}(1) \frac{(x-1)^k}{k!}.$$

On calcul alors les dérivées  $P^{(k)}(x)$  pour  $k = 0, \dots, 6$  et on les évalue en 1 :

$$\begin{array}{ll} P(x) = x^6 - 5x^5 + 10x^4 - 12x^3 + 11x^2 - 7x + 2 & P(1) = 0 \\ P'(x) = 6x^5 - 25x^4 + 40x^3 - 36x^2 + 22x - 7 & P'(1) = 0 \\ P''(x) = 30x^4 - 100x^3 + 120x^2 - 72x + 22 & P''(1) = 0 \\ P'''(x) = 120x^3 - 300x^2 + 240x - 72 & P'''(1) = -12 \\ P^{IV}(x) = 360x^2 - 600x + 240 & P^{IV}(1) = 0 \\ P^V(x) = 720x - 600 & P^V(1) = 120 \\ P^{VI}(x) = 720 & P^{VI}(1) = 720 \end{array}$$

La formule de Taylor en 1 s'écrit alors

$$\begin{aligned} P(x) &= -2(x-1)^3 + (x-1)^5 + (x-1)^6 = \\ &= (x-1)^3((x-1)^3 + (x-1)^2 - 2) = \\ &= (x-1)^3(x^3 - 3x^2 + 3x - 1 + x^2 - 2x + 1 - 2) = \\ &= (x-1)^3(x^3 - 2x^2 + x - 2) \end{aligned}$$

Ceci implique que 1 est une racine de  $P$  de multiplicité 3.

2. On pose  $Q(x) = x^3 - 2x^2 + x - 2$ . Une racine réelle évidente de  $Q$  est 2. En effectuant la division euclidienne de  $Q$  par  $(x-2)$  on obtient

$$Q(x) = (x-2)(x^2 + 1).$$

Étant donné que  $x^2 + 1$  ne se factorise pas sur  $\mathbb{R}$ , on conclut que la factorisation de  $P$  sur  $\mathbb{R}$  est

$$P(x) = (x-1)^3(x-2)(x^2 + 1).$$

Pour factoriser  $P$  sur  $\mathbb{C}$  il ne reste que factoriser le polynôme  $x^2 + 1$  sur  $\mathbb{C}$  ; on a les deux racines complexes conjuguées  $x_1 = -i$  et  $x_2 = +i$ . On conclut que la factorisation de  $P$  sur  $\mathbb{C}$  est

$$P(x) = (x-1)^3(x-2)(x-i)(x+i).$$

### Exercice 1.44

On considère le polynôme

$$P(x) = x^8 + 5x^7 + 9x^6 + 7x^5 + 3x^4 + 5x^3 + 9x^2 + 7x + 2.$$

1. Écrire la formule de Taylor de  $P$  en  $-1$  ; montrer que  $-1$  est une racine de  $P$  et déterminer son ordre de multiplicité.
2. Factoriser  $P$  d'abord sur  $\mathbb{R}$ , puis sur  $\mathbb{C}$  en sachant qu'il possède une autre racine réelle évidente (*i.e.* elle est un diviseur du terme de degré 0).

**Correction**

1. Pour un polynôme  $P$  de degré 8, la formule de Taylor en  $-1$  s'écrit

$$P(x) = \sum_{k=0}^{k=8} P^{(k)}(-1) \frac{(x+1)^k}{k!}.$$

On calcul alors les dérivées  $P^{(k)}(x)$  pour  $k = 0, \dots, 8$  et on les évalue en  $-1$  :

$$\begin{aligned} P(x) &= x^8 + 5x^7 + 9x^6 + 7x^5 + 3x^4 + 5x^3 + 9x^2 + 7x + 2 & P(-1) &= 0 \\ P'(x) &= 8x^7 + 35x^6 + 54x^5 + 35x^4 + 12x^3 + 15x^2 + 18x + 7 & P'(-1) &= 0 \\ P''(x) &= 56x^6 + 210x^5 + 270x^4 + 140x^3 + 36x^2 + 30x + 18 & P''(-1) &= 0 \\ P'''(x) &= 336x^5 + 1050x^4 + 1080x^3 + 420x^2 + 72x + 30 & P'''(-1) &= 12 \\ P^{IV}(x) &= 1680x^4 + 4200x^3 + 3240x^2 + 840x + 72 & P^{IV}(-1) &= -48 \\ P^V(x) &= 6720x^3 + 12600x^2 + 6480x + 840 & P^V(-1) &= 240 \\ P^{VI}(x) &= 20160x^2 + 25200x + 6480 & P^{VI}(-1) &= 1440 \\ P^{VII}(x) &= 40320x + 25200 & P^{VII}(-1) &= -15120 \\ P^{VIII}(x) &= 40320 & P^{VIII}(-1) &= 40320 \end{aligned}$$

La formule de Taylor en  $-1$  s'écrit alors

$$\begin{aligned} P(x) &= 2(x+1)^3 - 2(x+1)^4 + 2(x+1)^5 + 2(x+1)^6 - 3(x+1)^7 + (x+1)^8 = \\ &= (x+1)^3 [2 - 2(x+1) + 2(x+1)^2 + 2(x+1)^3 - 3(x+1)^4 + (x+1)^5] = \\ &= (x+1)^3 [2x^2 + 2x + 6 + 2(x^3 + 3x^2 + 3x + 1) - 3(x^4 + 4x^3 + 6x^2 + 4x + 1) + (x^5 + 5x^4 + 10x^3 + 10x^2 + 5x + 1)] = \\ &= (x+1)^3 (x^5 + 2x^4 + x + 2) \end{aligned}$$

Ceci implique que  $-1$  est une racine de  $P$  de multiplicité 3.

2. On pose  $Q(x) = x^5 + 2x^4 + x + 2$ . Une racine réelle évidente de  $Q$  est  $-2$ . En effectuant la division euclidienne de  $Q$  par  $(x+2)$  on obtient

$$Q(x) = (x+2)(x^4 + 1).$$

Étant donné que  $x^4 + 1$  ne se factorise pas sur  $\mathbb{R}$ , on conclut que la factorisation de  $P$  sur  $\mathbb{R}$  est

$$P(x) = (x+1)^3 (x+2)(x^4 + 1).$$

Pour factoriser  $P$  sur  $\mathbb{C}$  il ne reste que factoriser le polynôme  $x^4 + 1$  sur  $\mathbb{C}$ ; on a les quatre racines  $x_1 = \frac{1+i}{\sqrt{2}}$ ,  $x_2 = \frac{1-i}{\sqrt{2}}$ ,  $x_3 = \frac{-1+i}{\sqrt{2}}$  et  $x_4 = \frac{-1-i}{\sqrt{2}}$ . On conclut que la factorisation de  $P$  sur  $\mathbb{C}$  est

$$P(x) = (x+1)^3 (x+2) \left( x - \frac{1+i}{\sqrt{2}} \right) \left( x - \frac{1-i}{\sqrt{2}} \right) \left( x - \frac{-1+i}{\sqrt{2}} \right) \left( x - \frac{-1-i}{\sqrt{2}} \right).$$

**Exercice 1.45**

On considère le polynôme

$$P(x) = x^6 - x^5 - x^4 + x^2 + x - 1.$$

1. Écrire la formule de Taylor de  $P$  en 1 ; montrer que 1 est une racine de  $P$  et déterminer son ordre de multiplicité.
2. Factoriser  $P$  d'abord sur  $\mathbb{R}$ , puis sur  $\mathbb{C}$  en sachant qu'il possède une autre racine réelle évidente (*i.e.* elle est un diviseur du terme de degré 0).

**Correction**

1. Pour un polynôme  $P$  de degré 6, la formule de Taylor en 1 s'écrit

$$P(x) = \sum_{k=0}^{k=6} P^{(k)}(1) \frac{(x-1)^k}{k!}.$$

On calcul alors les dérivées  $P^{(k)}(x)$  pour  $k = 0, \dots, 6$  et on les évalue en 1 :

$$P(x) = x^6 - x^5 - x^4 + x^2 + x - 1 \quad P(1) = 0$$

$$\begin{aligned}
P'(x) &= 6x^5 - 5x^4 - 4x^3 + 2x + 1 & P'(1) &= 0 \\
P''(x) &= 30x^4 - 20x^3 - 12x^2 + 2 & P''(1) &= 0 \\
P'''(x) &= 120x^3 - 60x^2 - 24x & P'''(1) &= 36 \\
P^{IV}(x) &= 360x^2 - 120x - 24 & P^{IV}(1) &= 216 \\
P^V(x) &= 720x - 120 & P^V(1) &= 600 \\
P^{VI}(x) &= 720 & P^{VI}(1) &= 720
\end{aligned}$$

La formule de Taylor en 1 s'écrit alors

$$\begin{aligned}
P(x) &= 36 \frac{(x-1)^3}{6} + 216 \frac{(x-1)^4}{24} + 600 \frac{(x-1)^5}{120} + 720 \frac{(x-1)^6}{720} \\
&= 6(x-1)^3 + 9(x-1)^4 + 5(x-1)^5 + (x-1)^6 \\
&= (x-1)^3 \left( 6 + 9(x-1) + 5(x-1)^2 + (x-1)^3 \right) \\
&= (x-1)^3 \left( x^3 + 2x^2 + 2x + 1 \right)
\end{aligned}$$

Ceci implique que 1 est une racine de  $P$  de multiplicité 3.

2. On pose  $Q(x) = x^3 + 2x^2 + 2x + 1$ . Une racine réelle évidente de  $Q$  est  $-1$ . En effectuant la division euclidienne de  $Q$  par  $(x+1)$  on obtient

$$Q(x) = (x+1)(x^2 + x + 1).$$

Étant donné que  $x^2 + x + 1$  ne se factorise pas sur  $\mathbb{R}$ , on conclut que la factorisation de  $P$  sur  $\mathbb{R}$  est

$$P(x) = (x-1)^3(x+1)(x^2 + x + 1).$$

Pour factoriser  $P$  sur  $\mathbb{C}$  il ne reste que factoriser le polynôme  $x^2 + x + 1$  sur  $\mathbb{C}$ . Avec les notations usuelles du cours, on a  $\Delta = -3$  et une racine carrée de  $\Delta$  est  $\delta = i\sqrt{3}$  d'où les deux racines complexes conjuguées  $x_1 = -\frac{1}{2} + i\frac{\sqrt{3}}{2} = e^{i\frac{2}{3}\pi}$  et  $x_2 = -\frac{1}{2} - i\frac{\sqrt{3}}{2} = e^{i\frac{4}{3}\pi}$ . On conclut que la factorisation de  $P$  sur  $\mathbb{C}$  est

$$P(x) = (x-1)^3(x+1)(x - e^{i\frac{2}{3}\pi})(x - e^{i\frac{4}{3}\pi}).$$

### Exercice 1.46

On considère le polynôme

$$P(x) = x^6 + 3x^5 + 3x^4 - 3x^2 - 3x - 1.$$

1. Écrire la formule de Taylor de  $P$  en  $-1$ ; montrer que  $-1$  est une racine de  $P$  et déterminer son ordre de multiplicité.
2. Factoriser  $P$  d'abord sur  $\mathbb{R}$ , puis sur  $\mathbb{C}$  en sachant qu'il possède une autre racine réelle évidente (*i.e.* elle est un diviseur du terme de degré 0).

### Correction

1. Pour un polynôme  $P$  de degré 6, la formule de Taylor en  $-1$  s'écrit

$$P(x) = \sum_{k=0}^{k=6} P^{(k)}(-1) \frac{(x+1)^k}{k!}.$$

On calcul alors les dérivées  $P^{(k)}(x)$  pour  $k = 0, \dots, 6$  et on les évalue en  $-1$ :

$$\begin{aligned}
P(x) &= x^6 + 3x^5 + 3x^4 - 3x^2 - 3x - 1 & P(-1) &= 0 \\
P'(x) &= 6x^5 + 15x^4 + 12x^3 - 6x - 3 & P'(-1) &= 0 \\
P''(x) &= 30x^4 + 60x^3 + 36x^2 - 6 & P''(-1) &= 0 \\
P'''(x) &= 120x^3 + 180x^2 + 72x & P'''(-1) &= -12 \\
P^{IV}(x) &= 360x^2 + 360x + 72 & P^{IV}(-1) &= 72 \\
P^V(x) &= 720x + 360 & P^V(-1) &= -360 \\
P^{VI}(x) &= 720 & P^{VI}(-1) &= 720
\end{aligned}$$

La formule de Taylor en  $-1$  s'écrit alors

$$\begin{aligned} P(x) &= -12 \frac{(x+1)^3}{6} + 76 \frac{(x+1)^4}{24} - 360 \frac{(x+1)^5}{120} + 720 \frac{(x+1)^6}{720} \\ &= -2(x+1)^3 + 3(x+1)^4 - 3(x+1)^5 + (x+1)^6 \\ &= (x+1)^3(-2 + 3(x+1) - 3(x+1)^2 + (x+1)^3) \\ &= (x+1)^3(x^3 - 1) \end{aligned}$$

Ceci implique que  $-1$  est une racine de  $P$  de multiplicité 3.

2. On pose  $Q(x) = x^3 - 1$ . Une racine réelle évidente de  $Q$  est 1. En effectuant la division euclidienne de  $Q$  par  $(x-1)$  on obtient

$$Q(x) = (x-1)(x^2 + x + 1).$$

Étant donné que  $x^2 + x + 1$  ne se factorise pas sur  $\mathbb{R}$ , on conclut que la factorisation de  $P$  sur  $\mathbb{R}$  est

$$P(x) = (x+1)^3(x-1)(x^2 + x + 1).$$

Pour factoriser  $P$  sur  $\mathbb{C}$  il ne reste que factoriser le polynôme  $x^2 + x + 1$  sur  $\mathbb{C}$ . Avec les notations usuelles du cours, on a  $\Delta = -3$  et une racine carrée de  $\Delta$  est  $\delta = i\sqrt{3}$  d'où les deux racines complexes conjuguées  $x_1 = -\frac{1}{2} + i\frac{\sqrt{3}}{2} = e^{i\frac{2}{3}\pi}$  et  $x_2 = -\frac{1}{2} - i\frac{\sqrt{3}}{2} = e^{i\frac{4}{3}\pi}$ . On conclut que la factorisation de  $P$  sur  $\mathbb{C}$  est

$$P(x) = (x+1)^3(x-1)(x - e^{i\frac{2}{3}\pi})(x - e^{i\frac{4}{3}\pi}).$$

### Exercice 1.47

On considère le polynôme

$$P(x) = x^5 - x^3 - x^2 + 1.$$

1. Écrire la formule de Taylor de  $P$  en 1 ; montrer que 1 est une racine de  $P$  et déterminer son ordre de multiplicité.
2. Factoriser  $P$  d'abord sur  $\mathbb{R}$ , puis sur  $\mathbb{C}$  en sachant qu'il possède une autre racine réelle évidente (*i.e.* elle est un diviseur du terme de degré 0).

### Correction

1. Pour un polynôme  $P$  de degré 5, la formule de Taylor en 1 s'écrit

$$P(x) = \sum_{k=0}^{k=5} P^{(k)}(1) \frac{(x-1)^k}{k!}.$$

On calcul alors les dérivées  $P^{(k)}(x)$  pour  $k = 0, \dots, 5$  et on les évalue en 1 :

$$\begin{array}{ll} P(x) = x^5 - x^3 - x^2 + 1 & P(1) = 0 \\ P'(x) = 5x^4 - 3x^2 - 2x & P'(1) = 0 \\ P''(x) = 20x^3 - 6x - 2 & P''(1) = 12 \\ P'''(x) = 60x^2 - 6 & P'''(1) = 54 \\ P^{IV}(x) = 120x & P^{IV}(1) = 120 \\ P^V(x) = 120 & P^V(1) = 120 \end{array}$$

La formule de Taylor en 1 s'écrit alors

$$\begin{aligned} P(x) &= 6(x-1)^2 + 9(x-1)^3 + 5(x-1)^4 + (x-1)^5 = \\ &= (x-1)^2(6 + 9(x-1) + 5(x-1)^2 + (x-1)^3) = \\ &= (x-1)^2(6 + (9x-9) + (5x^2 - 10x + 5) + (x^3 - 3x^2 + 3x - 1)) = \\ &= (x-1)^2(x^3 + 2x^2 + 2x + 1) \end{aligned}$$

Ceci implique que 1 est une racine de  $P$  de multiplicité 2.

2. On pose  $Q(x) = x^3 + 2x^2 + 2x + 1$ . Une racine réelle évidente de  $Q$  est  $-1$ . En effectuant la division euclidienne de  $Q$

par  $(x+1)$  on obtient

$$Q(x) = (x+1)(x^2 + x + 1).$$

Étant donné que  $x^2 + x + 1$  ne se factorise pas sur  $\mathbb{R}$ , on conclut que la factorisation de  $P$  sur  $\mathbb{R}$  est

$$P(x) = (x-1)^2(x+1)(x^2 + x + 1).$$

Pour factoriser  $P$  sur  $\mathbb{C}$  il ne reste que factoriser le polynôme  $x^2 + x + 1$  sur  $\mathbb{C}$ ; on a les deux racines complexes conjuguées  $x_1 = (-1 + i\sqrt{3})/2$  et  $x_2 = (-1 - i\sqrt{3})/2$ . On conclut que la factorisation de  $P$  sur  $\mathbb{C}$  est

$$P(x) = (x-1)^2(x+1)\left(x - \frac{-1+i\sqrt{3}}{2}\right)\left(x - \frac{-1-i\sqrt{3}}{2}\right).$$

### Exercice 1.48

On considère le polynôme

$$P(x) = x^5 + 2x^4 + x^3 - x^2 - 2x - 1.$$

1. Écrire la formule de Taylor de  $P$  en  $-1$ ; montrer que  $-1$  est une racine de  $P$  et déterminer son ordre de multiplicité.
2. Factoriser  $P$  d'abord sur  $\mathbb{R}$ , puis sur  $\mathbb{C}$  en sachant qu'il possède une autre racine réelle évidente (*i.e.* elle est un diviseur du terme de degré 0).

#### Correction

1. Pour un polynôme  $P$  de degré 5, la formule de Taylor en  $-1$  s'écrit

$$P(x) = \sum_{k=0}^{k=5} P^{(k)}(-1) \frac{(x+1)^k}{k!}.$$

On calcul alors les dérivées  $P^{(k)}(x)$  pour  $k = 0, \dots, 5$  et on les évalue en 1 :

$$\begin{aligned} P(x) &= x^5 + 2x^4 + x^3 - x^2 - 2x - 1 & P(-1) &= 0 \\ P'(x) &= 5x^4 + 8x^3 + 3x^2 - 2x - 2 & P'(-1) &= 0 \\ P''(x) &= 20x^3 + 24x^2 + 6x - 2 & P''(-1) &= -4 \\ P'''(x) &= 60x^2 + 48x + 6 & P'''(-1) &= 18 \\ P^{IV}(x) &= 120x + 48 & P^{IV}(-1) &= -72 \\ P^V(x) &= 120 & P^V(-1) &= 120 \end{aligned}$$

La formule de Taylor en  $-1$  s'écrit alors

$$\begin{aligned} P(x) &= -2(x+1)^2 + 3(x+1)^3 - 3(x+1)^4 + (x+1)^5 = \\ &= (x+1)^2(-2 + 3(x+1) - 3(x+1)^2 + (x+1)^3) = \\ &= (x+1)^2(-2 + (3x+3) - (3x^2 + 6x + 3) + (x^3 + 3x^2 + 3x + 1)) = \\ &= (x+1)^2(x^3 - 1) \end{aligned}$$

Ceci implique que  $-1$  est une racine de  $P$  de multiplicité 2.

2. On pose  $Q(x) = x^3 - 1$ . Une racine réelle évidente de  $Q$  est 1. En effectuant la division euclidienne de  $Q$  par  $(x-1)$  on obtient

$$Q(x) = (x-1)(x^2 + x + 1).$$

Étant donné que  $x^2 + x + 1$  ne se factorise pas sur  $\mathbb{R}$ , on conclut que la factorisation de  $P$  sur  $\mathbb{R}$  est

$$P(x) = (x+1)^2(x-1)(x^2 + x + 1).$$

Pour factoriser  $P$  sur  $\mathbb{C}$  il ne reste que factoriser le polynôme  $x^2 + x + 1$  sur  $\mathbb{C}$ ; on a les deux racines complexes conjuguées  $x_1 = (-1 + i\sqrt{3})/2$  et  $x_2 = (-1 - i\sqrt{3})/2$ . On conclut que la factorisation de  $P$  sur  $\mathbb{C}$  est

$$P(x) = (x+1)^2(x-1)\left(x - \frac{-1+i\sqrt{3}}{2}\right)\left(x - \frac{-1-i\sqrt{3}}{2}\right).$$

### Exercice 1.49 (factorisation)

Factoriser les polynômes suivants en exploitant les informations données

1.  $P(x) = x^3 + 2x^2 - x - 2$ , (P possède une racine réelle),
2.  $P(x) = 2x^3 - (5+6i)x^2 + 9ix + 1 - 3i$  (P possède une racine réelle),
3.  $P(x) = x^3 - 2x^2 - (i+2)x + 3i - 3$  (P possède une racine réelle),
4.  $P(x) = x^3 + (3i-2)x^2 - (2+6i)x + 4$  (P possède une racine réelle),
5.  $P(x) = x^4 + \frac{1}{2}x^3 + ix + \frac{1}{2}i$  (P possède une racine réelle),
6.  $P(x) = x^5 + 3x^4 + 4x^3 + 4x^2 + 3x + 1$  (P possède une racine évidente),
7.  $P(x) = 2x^4 + x^3 - 6x^2 + x + 2$  (P possède une racine évidente).

#### Correction

1. Une racine évidente de  $P$  est  $x = 1$ . En effectuant la division euclidienne de  $P$  par  $(x - 1)$  on obtient

$$\begin{array}{r|rrrr} x^3 & +2x^2 & -x & -2 \\ \hline -x^3 & +x^2 & & \\ \hline 3x^2 & -x & -2 \\ -3x^2 & +3x & & \\ \hline 2x & -2 & & \\ -2x & +2 & & \\ \hline 0 & & & \end{array}$$

$$\text{donc } P(x) = (x-1)(x^2+3x+2) = (x-1)(x+1)(x+2).$$

2. Comme  $P(3) = 0$  on factorise  $P$  en effectuant la division euclidienne de  $P$  par  $(x - \frac{1}{2})$  :

$$\begin{array}{r|rrrr} 2x^3 & -(5+6i)x^2 & +9ix & +1-3i \\ \hline -2x^3 & +6x^2 & & \\ \hline (1-6i)x^2 & +9ix & +1-3i \\ -(1-6i)x^2 & +3(1-6i)x & & \\ \hline (3-9i)x & +1-3i \\ -(3-9i)x & -(3-9i) & & \\ \hline 0 & & & \end{array}$$

$$\text{On obtient } P(x) = (x-3)(2x^2 + (1-6i)x + (1-3i)).$$

3. Soit  $\alpha \in \mathbb{R}$  une racine réelle de  $P$ , alors elle satisfait l'équation  $2\alpha^3 - (5+6i)\alpha^2 + 9i\alpha + 1 - 3i = 0$ . En comparant les parties réelle et imaginaire dans cette égalité on trouve

$$\begin{cases} 2\alpha^3 - 5\alpha^2 + 1 = 0 \\ -6\alpha^2 + 9\alpha - 3 = 0. \end{cases}$$

La résolution de  $-6\alpha^2 + 9\alpha - 3 = 0$  donne deux solutions :  $\alpha = 1$  et  $\alpha = 1/2$ . Le réel  $\alpha = 1$  ne vérifiant pas l'équation  $2\alpha^3 - 5\alpha^2 + 1 = 0$  et le réel  $\alpha = 1/2$  oui, on a bien  $P(1/2) = 0$ . On factorise alors  $P$  en effectuant la division euclidienne de  $P$  par  $(x - 1/2)$  :

$$\begin{array}{r|rrrr} 2x^3 & -(5+6i)x^2 & +9ix & +1-3i \\ \hline -2x^3 & +x^2 & & \\ \hline -(4+6i)x^2 & +9ix & +1-3i \\ -(4+6i)x^2 & -(2+3i)x & & \\ \hline -(2-6i)x & 1-3i \\ -(2-6i)x & -(1-3i) & & \\ \hline 0 & & & \end{array}$$

$$\text{On obtient } P(x) = (x-1/2)(2x^2 - (4+6i)x - (2-6i)). \text{ On va maintenant factoriser } 2x^2 - (4+6i)x - (2-6i) :$$

$$2x^2 - (4+6i)x - (2-6i) = 0 \implies \delta^2 = ((4+6i))^2 - 4(2)(-(2-6i)) = -4 \implies \delta = 2i \implies x_0 = 1+2i, x_1 = 1+i;$$

$$\text{par conséquent } P(x) = 2(x-1/2)(x-1-2i)(x-1-i).$$

4. Comme  $P(2) = 0$  on factorise  $P$  en effectuant la division euclidienne de  $P$  par  $(x - 2)$  :

$$\begin{array}{r} x^3 + (3i-2)x^2 - (6i+2)x + 4 \\ \hline -x^3 + 2x^2 \\ \hline 3ix^2 - (6i+2)x + 4 \\ -3ix^2 + 6ix \\ \hline -2x + 4 \\ 2x - 4 \\ \hline 0 \end{array} \quad \left| \begin{array}{r} x - 2 \\ x^2 + 3ix - 2 \end{array} \right.$$

On obtient  $P(x) = (x - 2)(x^2 + 3ix - 2)$ . On voit que  $x = -i$  est une racine du polynôme  $x^2 + 3ix - 2$  et on trouve  $P(x) = (x - 2)(x + i)(x + 2i)$ .

5. Soit  $\alpha \in \mathbb{R}$  une racine réelle de  $P$ , alors elle satisfait l'équation  $2\alpha^4 + \frac{1}{2}\alpha^3 + i\alpha + \frac{1}{2}i = 0$ . En comparant les parties réelle et imaginaire dans cette égalité on trouve  $\alpha = -1/2$  et on a bien  $P(-1/2) = 0$ . On factorise alors  $P$  en effectuant la division euclidienne de  $P$  par  $(x + 1/2)$  :

$$\begin{array}{r} x^4 + \frac{1}{2}x^3 + ix + \frac{1}{2}i \\ \hline -x^4 - \frac{1}{2}x^3 \\ \hline +ix + \frac{1}{2}i \\ -ix - \frac{1}{2}i \\ \hline 0 \end{array} \quad \left| \begin{array}{r} x + 1/2 \\ x^3 + i \end{array} \right.$$

On obtient  $P(x) = (x - 1/2)(x^3 + i)$ . On remarque que  $x^3 + i = x^3 - i^3 = (x - i)(x^2 + ix - 1)$ . On va maintenant factoriser  $x^2 + ix - 1$  :

$$x^2 + ix - 1 = 0 \implies \delta^2 = 3 \implies \delta = \sqrt{3} \implies x_0 = \frac{\sqrt{3}-i}{2}, x_1 = \frac{-\sqrt{3}-i}{2};$$

par conséquent  $P(x) = (x - 1/2)(x - i)\left(x - \frac{\sqrt{3}-i}{2}\right)\left(x - \frac{-\sqrt{3}-i}{2}\right)$ .

6. Comme  $P(-1) = 0$  on sait que  $P$  est divisible par  $(x + 1)$ . Avant d'effectuer la division euclidienne, cherchons d'abord la multiplicité de cette racine :

$$\begin{array}{ll} P(x) = x^5 + 3x^4 + 4x^3 + 4x^2 + 3x + 1 & P(-1) = 0 \\ P'(x) = 5x^4 + 12x^3 + 12x^2 + 8x + 3 & P'(-1) = 0 \\ P''(x) = 20x^3 + 36x^2 + 24x + 8 & P''(-1) = 0 \\ P'''(x) = 60x^2 + 72x + 24 & P'''(-1) = 12 \end{array}$$

On sait alors que  $P(x) = (x + 1)^3 Q(x)$  et on calcule  $Q$  en effectuant la division euclidienne de  $P$  par  $(x + 1)^3$  :

$$\begin{array}{r} x^5 + 3x^4 + 4x^3 + 4x^2 + 3x + 1 \\ \hline -x^5 - 3x^4 - 3x^3 - x^2 \\ \hline x^3 + 3x^2 + 3x + 1 \\ -x^3 - 3x^2 - 3x - 1 \\ \hline 0 \end{array} \quad \left| \begin{array}{r} x^3 + 3x^2 + 3x + 1 \\ x^2 + 1 \end{array} \right.$$

donc  $P(x) = (x + 1)^3(x^2 + 1) = (x + 1)^3(x - i)(x + i)$ .

7. Comme  $P(1) = 0$  on sait que  $P$  est divisible par  $(x - 1)$ . Avant d'effectuer la division euclidienne, cherchons d'abord la multiplicité de cette racine :

$$\begin{array}{ll} P(x) = 2x^4 + x^3 - 6x^2 + x + 2 & P(1) = 0 \\ P'(x) = 8x^3 + 3x^2 - 12x + 1 & P'(1) = 0 \\ P''(x) = 24x^2 + 6x - 12 & P''(1) = 30 \end{array}$$

On sait alors que  $P(x) = (x-1)^2 Q(x)$  et on calcule  $Q$  en effectuant la division euclidienne de  $P$  par  $(x-1)^2$  :

$$\begin{array}{r} 2x^4 & +x^3 & -6x^2 & +x & +2 \\ \hline -2x^4 & +4x^3 & -2x^2 & & \\ \hline 5x^3 & -8x^2 & +x & +2 \\ -5x^3 & +10x^2 & -5x & & \\ \hline 2x^2 & +4x & -2 \\ -2x^2 & -4x & +2 \\ \hline 0 \end{array} \quad \left| \begin{array}{l} x^2 - 2x + 1 \\ 2x^2 + 5x + 2 \end{array} \right.$$

On obtient  $P(x) = (x-1)^2(2x^2 + 5x + 2) = 2(x-1)^2(x+1/2)(x+2)$ .

### Exercice 1.50 (multiplicité)

Montrer que pour tout  $n \in \mathbb{N}^*$  le polynôme  $P_n(x) := 1 + \frac{x}{1!} + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!}$  n'a pas de racines multiples.  
On pourra calculer  $P'_n$  et l'exprimer en fonction de  $P_n$ .

#### Correction

- ①  $P_n(x) = \sum_{k=0}^n \frac{x^k}{k!}$  (rappel :  $x^0 = 1$  et  $0! = 1$ )
  - ②  $\alpha = 0$  n'est pas racine de  $P_n(x)$  car  $P_n(0) = 1$
  - ③  $P'_n(x) = \sum_{k=0}^n k \frac{x^{k-1}}{k!} = \sum_{j=0}^{n-1} (j+1) \frac{x^j}{(j+1)!} = P_{n-1}(x)$
  - ④  $P_n(x) = \sum_{k=0}^n \frac{x^k}{k!} = \sum_{k=0}^{n-1} \frac{x^k}{k!} + \frac{x^n}{n!} = P'_n(x) + \frac{x^n}{n!}$
- Si  $\alpha$  est racine de  $P_n$  alors  $\alpha \neq 0$  et  $P'_n(\alpha) = -\frac{\alpha^n}{n!} \neq 0$  donc  $\alpha$  est une racine simple de  $P_n$ .

### Exercice 1.51

Soient  $a$  et  $b$  deux nombres complexes et  $P(x) := x^4 - 2ix^3 + 3(1+i)x^2 + ax + b$ .

- ★ Calculer  $a$  et  $b$  pour que  $i$  soit une racine multiple de  $P$ .

Dans toute la suite, on suppose que  $i$  est une racine multiple de  $P$ .

- ★ Quelle est la multiplicité de la racine  $i$  de  $P$  ?
- ★ Calculer les autres racines de  $P$ .

#### Correction

- ★ Pour que  $i$  soit une racine multiple de  $P$  il faut que  $P(i) = 0$  et  $P'(i) = 0$  :

$$P(x) := x^4 - 2ix^3 + 3(1+i)x^2 + ax + b \quad P(i) = -4 + 3i + ai + b$$

donc  $P(i) = 0$  si et seulement si  $b = 4 + 3i - ai$  ;

$$P'(x) := 4x^3 - 6ix^2 + 6(1+i)x + a \quad P'(i) = 8i - 6 + a$$

donc  $P'(i) = 0$  si et seulement si  $a = 6 - 8i$  ce qui implique  $b = -4 - 3i$ .

Dans toute la suite, on suppose que  $a = 6 - 8i$  ce qui implique  $b = -4 - 3i$ .

- ★ On sait déjà que  $i$  a au moins multiplicité 2. Vérifions si c'est supérieure :

$$P''(x) = 12x^2 - 12ix + 6(1+i) \quad P''(i) = 6(1+i) \neq 0$$

donc elle a multiplicité 2.

- ★ On doit alors calculer deux autres racines de  $P$  : soit on effectue la division euclidienne de  $P$  par  $(x-i)^2$ , soit on factorise  $P$  à l'aide de la formule de Taylor, ce qui donne

$$\begin{aligned} P(x) &= P(i) + (x-i)P'(i) + \frac{(x-i)^2}{2!}P''(i) + \frac{(x-i)^3}{3!}P'''(i) + \frac{(x-i)^4}{4!}P^{IV}(i) \\ &= \frac{(x-i)^2}{2!}6(1+i) + \frac{(x-i)^3}{3!}12i + \frac{(x-i)^4}{4!}24 \\ &= (x-i)^2(x^2 + 4 + 3i) = (x-i)^2 \left( x - \frac{1-3i}{\sqrt{2}} \right) \left( x - \frac{1+3i}{\sqrt{2}} \right). \end{aligned}$$

**Exercice 1.52**

Montrer que le polynôme  $P(x) := 2x^4 - 7x^3 + 9x^2 - 5x + 1$  admet une racine triple que l'on déterminera, puis factoriser  $P$  à l'aide de la formule de Taylor.

**Correction**

Comme  $P(1) = 0$  on sait que  $P$  est divisible par  $(x - 1)$ . Cherchons la multiplicité de cette racine :

$$\begin{array}{ll} P(x) = 2x^4 - 7x^3 + 9x^2 - 5x + 1 & P(1) = 0 \\ P'(x) = 8x^3 - 21x^2 + 18x - 5 & P'(1) = 0 \\ P''(x) = 24x^2 - 42x + 18 & P''(1) = 0 \\ P'''(x) = 48x - 42 & P'''(1) = 6 \\ P^{IV}(x) = 48 & P^{IV}(1) = 48 \end{array}$$

On sait alors que  $P(x) = (x - 1)^3 Q(x)$  et la formule de Taylor en  $x = 1$  s'écrit

$$P(x) = 6 \frac{(x-1)^3}{3!} + 48 \frac{(x-1)^4}{4!} = \frac{(x-1)^3}{3!} \left( 6 + 48 \frac{(x-1)}{4} \right) = (x-1)^3 (2x-1).$$

**Exercice 1.53**

Factoriser  $P(x) := x^8 + x^4 + 1$  sur  $\mathbb{R}$  et sur  $\mathbb{C}$ .

*On remarquera pour commencer que  $P(x) = (x^4 + 1)^2 - x^4$ .*

**Correction**

$$\begin{aligned} P(x) &= x^8 + x^4 + 1 \\ &= (x^4 + 1)^2 - x^4 \\ &= [(x^4 + 1) - x^2][(x^4 + 1) + x^2] \\ &= [x^4 - x^2 + 1][x^4 + x^2 + 1] \\ &= [(x^2 - 1)^2 - x^2][(x^2 + 1)^2 - x^2] \\ &= [(x^2 - 1) - x][(x^2 - 1) + x][(x^2 + 1) - x][(x^2 + 1) + x] \\ &= (x^2 - x - 1)(x^2 + x - 1)(x^2 - x + 1)(x^2 + x + 1) \\ &= \left(x - \frac{1+\sqrt{3}}{2}\right)\left(x - \frac{1-\sqrt{3}}{2}\right)\left(x - \frac{-1+\sqrt{3}}{2}\right)\left(x - \frac{-1-\sqrt{3}}{2}\right)\left(x - \frac{1-i\sqrt{3}}{2}\right)\left(x - \frac{-1+i\sqrt{3}}{2}\right)\left(x - \frac{-1-i\sqrt{3}}{2}\right). \end{aligned}$$

**★ Exercice 1.54 (Représentation et manipulation de polynômes)**

Dans cette exercice nous allons construire des fonctions qui se trouvent déjà dans Octave, on pourra comparer donc le résultat obtenu avec celui d'Octave. Attention, vous devez programmer vous-même les fonctions indiquées. Toute utilisation de fonctions toutes prêtes ne sera pas prise en compte.

Soit  $\mathbb{R}_n[x]$  l'ensemble des polynômes de degré inférieur ou égale à  $n$ ,  $n \in \mathbb{N}^*$ . Tout polynôme de cet espace vectoriel s'écrit de manière unique comme

$$p_n(x) = \sum_{i=0}^n a_i x^i = a_0 + a_1 x + \cdots + a_n x^n, \quad \text{où } a_i \in \mathbb{R} \text{ pour } i = 0, \dots, n.$$

Les  $n+1$  valeurs réels  $a_0, a_1, \dots, a_n$  sont appelés les **coordonnées de  $p_n$  dans la base canonique**<sup>8</sup> de  $\mathbb{R}_n[x]$  et on peut les stocker dans un vecteur  $\mathbf{p}$  :

$$\mathbf{p} = \text{coord}(p_n, \mathcal{C}_n) = (a_0, a_1, a_2, \dots, a_n) \in \mathbb{R}^{n+1}$$

Dans Octave nous utiliserons le vecteur  $\mathbf{p}$  pour manipuler un polynôme et nous construirons des fonctions pour opérer sur les polynômes à partir de cette représentation. Par exemple, pour construire le polynôme  $p_2(x) = 2 - x + x^2$  nous écrirons

```
p=[2 -1 1]
```

Dans le **script** `script_pol.m` on écrira les instructions utilisées pour tester les fonction suivantes :

8. La base canonique de l'espace vectoriel  $\mathbb{R}_n[x]$  est l'ensemble  $\mathcal{C}_n = \{1, x, x^2, \dots, x^n\}$

1. Implémenter une fonction appelée `eval_pol` permettant d'évaluer le polynôme  $p$  (la fonction polynomiale) en des points donnés. La syntaxe doit être `function y=eval_pol(p,x)` où  $x$  est une valeur numérique ou un vecteur. Dans le second cas on doit obtenir un vecteur contenant les valeurs de la fonction polynomiale aux différents points spécifiés dans le vecteur  $\mathbf{x}$ . Par exemple, pour évaluer le polynôme  $p(x) = 1 + 2x + 3x^2$  en  $\mathbf{x} = (-1, 0, 1, 2)$  nous écrirons

```
p=[1 2 3]
y=eval_pol(p, [-1,0,1,2])
```

et on veut obtenir le vecteur  $\mathbf{y} = p(\mathbf{x}) = (2, 1, 6, 17)$ . En effet on a

$$\begin{aligned} p(-1) &= 1 + 2 \times (-1) + 3 \times ((-1)^2) = 1 - 2 + 3 = 2 \\ p(0) &= 1 + 2 \times 0 + 3 \times (0^2) = 1 + 0 + 0 = 1 \\ \mathbf{p} = \text{coord}(p, \mathcal{C}_2) &= (1, 2, 3) \quad p(1) = 1 + 2 \times 1 + 3 \times (1^2) = 1 + 2 + 3 = 6 \\ p(2) &= 1 + 2 \times 2 + 3 \times (2^2) = 1 + 4 + 12 = 17 \end{aligned}$$

2. Implémenter une fonction appelée `plot_pol` prenant en entrée un polynôme  $p$  (*i.e.* le vecteur qui contient ses coordonnées) et deux réels  $a$  et  $b > a$  et qui trace le graphe de  $p$  pour  $x \in [a, b]$ . La syntaxe de l'instruction doit être `plot_pol(p, a, b)`. Par exemple, pour tracer le graphe du polynôme  $p(x) = 1 + 2x + 3x^2$  sur l'intervalle  $[-2; 2]$  nous écrirons

```
p=[1 2 3]
plot_pol(p, -2, 2)
```

3. Implémenter une fonction appelée `sum_pol` renvoyant la somme de deux polynômes (attention, si les deux polynômes n'ont pas même degré, il faudra ajouter des zéros en fin du polynôme de plus petit degré afin de pouvoir calculer l'addition des deux vecteurs représentatifs). Par exemple, pour  $\mathbf{p} = (1, 2, 3)$  et  $\mathbf{q} = (1, -2)$ , on veut obtenir  $\mathbf{s} = (2, 0, 3)$  :

$$\begin{aligned} p(x) &= 1 + 2x + 3x^2 & \mathbf{p} = \text{coord}(p, \mathcal{C}_2) &= (1, 2, 3) \\ q(x) &= 1 - 2x & \mathbf{q} = \text{coord}(q, \mathcal{C}_1) &= (1, -2) \implies \mathbf{q} = \text{coord}(q, \mathcal{C}_2) &= (1, -2, 0) \\ s(x) &= p(x) + q(x) = 2 + 3x^2 & \mathbf{s} = \text{coord}(p + q, \mathcal{C}_2) &= (2, 0, 3) \end{aligned}$$

4. Implémenter une fonction appelée `prod_pol` renvoyant le produit de deux polynômes.

Exemple, pour  $\mathbf{p} = (1, 0, 3)$  et  $\mathbf{q} = (1, -2)$ , on veut obtenir  $\mathbf{u} = (1, -2, 3, -6)$ .

$$\begin{aligned} p(x) &= 1 + 3x^2 & \mathbf{p} = \text{coord}(p, \mathcal{C}_2) &= (1, 0, 3) \\ q(x) &= 1 - 2x & \mathbf{q} = \text{coord}(q, \mathcal{C}_2) &= (1, -2, 0) \\ u(x) &= p(x) \times q(x) = 1 \times p(x) - 2x \times p(x) = 1 - 2x + 3x^2 - 6x^3 & \mathbf{u} = \text{coord}(p \times q, \mathcal{C}_3) &= (1, -2, 3, -6) \end{aligned}$$

5. Implémenter une fonction appelée `derivee_pol` renvoyant la dérivée  $d$  du polynôme  $p$  donné en entrée (attention, si  $p \in \mathbb{R}_n[x]$ , alors  $d \in \mathbb{R}_{n-1}[x]$ ).

Exemple, pour  $\mathbf{p} = (1, 2, 6)$ , on veut obtenir  $\mathbf{d} = (2, 12)$ .

$$\begin{aligned} p(x) &= 1 + 2x + 6x^2 & \mathbf{p} = \text{coord}(p, \mathcal{C}_2) &= (1, 2, 6) \\ d(x) &= p'(x) = 2 + 12x & \mathbf{d} = \text{coord}(d, \mathcal{C}_1) &= (2, 12) \end{aligned}$$

6. Implémenter une fonction appelée `primitive_pol` renvoyant la primitive  $v$  du polynôme  $p$  donné en entrée ayant 0 pour racine (attention, si  $p \in \mathbb{R}_n[x]$ , alors  $v \in \mathbb{R}_{n+1}[x]$ ).

Exemple, pour  $\mathbf{p} = (1, 2, 6)$ , on veut obtenir  $\mathbf{v} = (0, 1, 1, 2)$ .

$$\begin{aligned} p(x) &= 1 + 2x + 6x^2 & \mathbf{p} = \text{coord}(p, \mathcal{C}_2) &= (1, 2, 6) \\ v(x) &= \int_0^x p(t) dt = \int_0^x 1 + 2t + 6t^2 dt = x + x^2 + 2x^3 & \mathbf{v} = \text{coord}(v, \mathcal{C}_3) &= (0, 1, 1, 2) \end{aligned}$$

7. Implémenter une fonction appelée `integrale_pol` renvoyant l'intégrale d'un polynôme entre deux valeurs  $a$  et  $b$ .

Exemple, pour  $\mathbf{p} = (1, 2, 6)$ ,  $a = 1$  et  $b = 2$ , on veut obtenir  $c = 18$  :

$$\begin{aligned} p(x) &= 1 + 2x + 6x^2 \\ c &= \int_a^b p(t) dt = \int_0^b p(t) dt - \int_0^a p(t) dt = v(b) - v(a) = b + b^2 + 2b^3 - a - a^2 - 2a^3 = 18. \end{aligned}$$

8. Implémenter une fonction appelée `print_pol` prenant en entrée un polynôme  $p$  (*i.e.* le vecteur qui contient ses coordonnées) et qui écrit dans la fenêtre de commande le polynôme dans la base canonique.

Exemple, pour  $\mathbf{p} = (1, 2, -3, 0, 7)$ , on veut afficher le message  $1+2x-3x^2+7x^4$ .

### Correction

Dans le fichier `script_pol.m` on écrit les instructions qui permettent de tester les différents points de cet exercice.

1. Dans le fichier `eval_pol` on écrit

```
function [y]=eval_pol(p,x)
y=zeros(size(x));
for k=1:length(p)
    y+=p(k)*x.^^(k-1);
end
end
```

et on teste cette fonction par exemple comme suit

```
y=eval_pol([1 2 3],[-1 0 1 2])
```

2. Dans le fichier `plot_pol.m` on écrit

```
function plot_pol(p,a,b)
x=linspace(a,b,100);
y=eval_pol(p,x);
plot(x,y);
end
```

et on teste cette fonction par exemple comme suit

```
plot_pol([-1 0 1],-2,2)
```

3. Sans perte de généralité, supposons que  $n > m$ , alors

$$\begin{aligned} p(x) &= \sum_{i=0}^n a_i x^i = \sum_{i=0}^m a_i x^i + \sum_{i=m+1}^n a_i x^i & \text{coord}(p, \mathcal{C}) = (a_0, a_1, a_2, \dots, a_m, a_{m+1}, \dots, a_n) \\ q(x) &= \sum_{i=0}^m b_i x^i = \sum_{i=0}^m b_i x^i + \sum_{i=m+1}^n 0 \times x^i & \text{coord}(q, \mathcal{C}) = (b_0, b_1, b_2, \dots, b_m) \\ (p+q)(x) &= \sum_{i=0}^m (a_i + b_i) x^i + \sum_{i=m+1}^n a_i x^i & \text{coord}(p+q, \mathcal{C}) = (a_0 + b_0, a_1 + b_1, a_2 + b_2, \dots, a_m + b_m, a_{m+1}, \dots, a_n) \end{aligned}$$

Dans le fichier `sum_pol.m` on écrit

```
function s=sum_pol(p,q)
n=length(p);
m=length(q);
A=zeros(2,max(n,m));
A(1,1:n)=p;
A(2,1:m)=q;
s=sum(A);
end
```

et on teste cette fonction par exemple comme suit

```
s=sum_pol([1 2 3],[4 5 6])
s=sum_pol([1 2 3],[4 5])
s=sum_pol([1 2],[4 5 6])
```

4. Dans le fichier `prod_pol.m` on écrit

```
function s=prod_pol(p,q)
n=length(p);
m=length(q);
A=zeros(m,n+m-1);
for i=1:m
    A(i,i:n+i-1)=q(i)*p;
end
s=sum(A);
end
```

et on teste cette fonction par exemple comme suit

```
u=prod_pol([1],[4 5 6])
u=prod_pol([1 2],[4 5 6])
u=prod_pol([1 2 3],[4 5 6])
u=prod_pol([1 2 3 4],[4 5 6])
```

5. Remarquons que

$$p(x) = \sum_{i=0}^n a_i x^i$$

$$\text{coord}(p, \mathcal{C}_n) = (a_0, a_1, a_2, \dots, a_n)$$

$$d(x) = p'(x) = \sum_{i=0}^n i a_i x^{i-1}$$

$$\text{coord}(d, \mathcal{C}_{n-1}) = (a_1, 2a_2, \dots, n a_n)$$

Dans le fichier *derivee\_pol.m* on écrit

```
function d=derivee_pol(p)
n=length(p);
d=p(2:end).*(1:n-1);
end
```

et on teste cette fonction par exemple comme suit

```
d=derivee_pol([1])
d=derivee_pol([1 2])
d=derivee_pol([1 2 3])
d=derivee_pol([1 2 1 1])
```

6. Remarquons que

$$p(x) = \sum_{i=0}^n a_i x^i$$

$$\text{coord}(p, \mathcal{C}_n) = (a_0, a_1, a_2, \dots, a_n)$$

$$v(x) = \int_0^x p(t) dt = \sum_{i=0}^n a_i \int_0^x t^i dt = \sum_{i=0}^n a_i \frac{x^{i+1}}{i+1}$$

$$\text{coord}(v, \mathcal{C}_{n+1}) = \left(0, \frac{a_0}{0+1}, \frac{a_1}{1+1}, \frac{a_2}{2+1}, \dots, \frac{a_n}{n+1}\right)$$

Dans le fichier *primitive\_pol.m* on écrit

```
function prim=primitive_pol(p)
n=length(p);
prim(1)=0;
prim([2:n+1])=p([1:n])./[1:n];
end
```

et on teste cette fonction par exemple comme suit

```
v=primitive_pol([1])
v=primitive_pol([1 2])
v=primitive_pol([1 2 3])
v=primitive_pol([1 2 1 1])
```

7. Dans le fichier *integrale\_pol.m* on écrit

```
function integr=integrale_pol(p,a,b)
prim=primitive_pol(p);
n=length(prim); % = 1+length(p)
aa([1:n])=a.^([0:n-1]);
prima=sum(prim.*aa);
bb([1:n])=b.^([0:n-1]);
primb=sum(prim.*bb);
integr=primb-prima;
end
```

et on teste cette fonction par exemple comme suit

```
w=integrale_pol([1 1], 1, 2)
```

8. Dans le fichier *print\_pol.m* on écrit

```
function str=print_pol(p)
n=length(p);
str='';
if n==1;
    str=strcat(num2str(p(1)));
else
    strsign=char((p>0)*'+' + (p<0)*'-' + (p==0)*'0');
    if p(1)~=0
        str=num2str(p(1));
    end
    if p(2)~=0
        str=strcat(str,strsign(2),num2str(p(2)),
                  'x');
    end
    for i=3:n
        if p(i)~=0
            str=strcat(str,strsign(i),num2str(p(i)),
                        'x',num2str(i-1));
        end
    end
end
end
```

et on teste cette fonction par exemple comme suit

```
print_pol([1 2 -3 -7 5])
print_pol([1 0 -3 0 5])
print_pol([1 2 -3 7])
print_pol([1 2 -3])
print_pol([1 2])
print_pol([1])
```

# Chapitre 2

## Interpolation

Étant donné  $n + 1$  couples  $\{(x_i, y_i)\}_{i=0}^n$ , le problème consiste à trouver une fonction  $\varphi = \varphi(x)$  telle que  $\varphi(x_i) = y_i$ ; on dit alors que  $\varphi$  interpole l'ensemble de valeurs  $\{y_i\}_{i=0}^n$  aux noeuds  $\{x_i\}_{i=0}^n$ . Les quantités  $y_i$  représentent les valeurs aux noeuds  $x_i$  d'une fonction  $f$  connue analytiquement ou des données expérimentales. Dans le premier cas, l'approximation a pour but de remplacer  $f$  par une fonction plus simple en vue d'un calcul numérique d'intégrale ou de dérivée. Dans l'autre cas, le but est d'avoir une représentation synthétique de données expérimentales (dont le nombre peut être très élevé). On parle d'*interpolation polynomiale* quand  $\varphi$  est un polynôme et d'*interpolation polynomiale par morceaux* (ou d'*interpolation par fonctions splines*) si  $\varphi$  est polynomiale par morceaux.

### 2.1 Interpolation polynomiale : base canonique, base de Lagrange, base de Newton

Supposons que l'on veuille chercher un polynôme  $p_n$  de degré  $n \geq 0$  qui, pour des valeurs  $x_0, x_1, x_2, \dots, x_n$  distinctes données (appelés noeuds d'interpolation), prenne les valeurs  $y_0, y_1, y_2, \dots, y_n$  respectivement, c'est-à-dire

$$p_n(x_i) = y_i \quad \text{pour } 0 \leq i \leq n. \quad (2.1)$$

Si un tel polynôme existe, il est appelé *polynôme d'interpolation* ou *polynôme interpolant*.

**Base canonique.** Une manière apparemment simple de résoudre ce problème est d'écrire le polynôme dans la base canonique de  $\mathbb{R}_n[x]$ :

$$p_n(x) = a_0 + a_1 x + a_2 x^2 + \dots + a_n x^n,$$

où  $a_0, a_1, a_2, \dots, a_n$  sont des coefficients qui devront être déterminés. Les  $(n + 1)$  relations (2.1) s'écrivent alors

$$\begin{cases} a_0 + a_1 x_0 + \dots + a_n x_0^n = y_0 \\ a_0 + a_1 x_1 + \dots + a_n x_1^n = y_1 \\ \dots \\ a_0 + a_1 x_n + \dots + a_n x_n^n = y_n \end{cases}$$

Puisque les valeurs  $x_i$  et  $y_i$  sont connues, ces relations forment un système linéaire de  $(n + 1)$  équations en les  $(n + 1)$  inconnues  $a_0, a_1, a_2, \dots, a_n$  qu'on peut mettre sous la forme matricielle

$$\begin{pmatrix} 1 & x_0 & \dots & x_0^n \\ 1 & x_1 & \dots & x_1^n \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \dots & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}. \quad (2.2)$$

Ainsi, le problème consistant à chercher le polynôme  $p_n$  satisfaisant (2.1) peut se réduire à résoudre le système linéaire (2.2) (cette matrice s'appelle matrice de VANDERMONDE).

Étant donné  $n + 1$  points distincts  $x_0, \dots, x_n$  et  $n + 1$  valeurs correspondantes  $y_0, \dots, y_n$ , il existe un unique polynôme  $p_n \in \mathbb{R}_n[x]$  tel que  $p_n(x_i) = y_i$ , pour  $i = 0, \dots, n$  qu'on peut écrire sous la forme

$$p_n(x) = \sum_{i=0}^n a_i x^i \quad \text{avec} \quad \begin{pmatrix} 1 & x_0 & \dots & x_0^n \\ 1 & x_1 & \dots & x_1^n \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \dots & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

**Base de Lagrange.** Malheureusement, résoudre une système linéaire de  $(n + 1)$  équations à  $(n + 1)$  inconnues n'est pas une tache triviale. Cette méthode pour trouver le polynôme  $p_n$  n'est donc pas une bonne méthode en pratique. On se

demande alors s'il existe une autre base  $\{L_0, L_1, L_2, \dots, L_n\}$  de  $\mathbb{R}_n[x]$  telle que le polynôme  $p_n$  s'écrit

$$p_n(x) = y_0 L_0(x) + y_1 L_1(x) + y_2 L_2(x) + \cdots + y_n L_n(x),$$

autrement dit s'il existe une base telle que les coordonnées du polynôme dans cette base ne sont rien d'autre que les valeurs connues  $y_0, y_1, \dots, y_n$ . Pour trouver une telle base, commençons par imposer le passage du polynôme par les  $n+1$  points donnés : les  $(n+1)$  relations (2.1) imposent la condition

$$L_i(x_j) = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{sinon} \end{cases} \quad \text{pour } 0 \leq i, j \leq n,$$

ce qui donne

$$L_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^n \frac{x - x_j}{x_i - x_j} = \frac{(x - x_0)(x - x_1) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n)}{(x_i - x_0)(x_i - x_1) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n)}.$$

Il est facile de vérifier que

- $L_i(x) \in \mathbb{R}_n[x]$  car le numérateur de  $L_i(x)$  est un produit de  $n$  termes  $(x - x_j)$  avec  $i \neq j$  et est donc un polynôme de degré  $n$  et le dénominateur de  $L_i(x)$  est une constante,
- $L_i(x_j) = 0$  si  $i \neq j$ ,  $0 \leq i \leq n$ ,
- $L_i(x_i) = 1$ .

De plus, les polynômes  $L_0, L_1, L_2, \dots, L_n$  sont linéairement indépendants car si l'équation  $\sum_{i=0}^n \alpha_i L_i(x) = 0$  doit être satisfait pour tout  $x \in \mathbb{R}$  alors en particulier elle doit être satisfait pour  $x = x_j$  pour tout  $j = 0, 1, \dots, n$  et puisque  $\sum_{i=0}^n \alpha_i L_i(x_j) = \alpha_j$ , on conclut que tous les  $\alpha_j$  sont nuls. Par conséquent, la famille  $\{L_0, L_1, L_2, \dots, L_n\}$  forme une base de  $\mathbb{R}_n[x]$ .

Il est important de remarquer que nous avons construit explicitement une solution du problème (2.1) et ceci pour n'importe quelles valeurs  $y_0, y_1, y_2, \dots, y_n$  données. Ceci montre que le système linéaire (2.2) a toujours une unique solution.

Étant donné  $n+1$  points distincts  $x_0, \dots, x_n$  et  $n+1$  valeurs correspondantes  $y_0, \dots, y_n$ , il existe un unique polynôme  $p_n \in \mathbb{R}_n[x]$  tel que  $p_n(x_i) = y_i$ , pour  $i = 0, \dots, n$  qu'on peut écrire sous la forme

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

Cette relation est appelée formule d'interpolation de LAGRANGE et les polynômes  $L_i$  sont les polynômes caractéristiques (de LAGRANGE).

**Base de Newton.** Cependant, cette méthode n'est pas encore la plus efficace d'un point de vue pratique. En effet, pour calculer le polynôme d'interpolation d'un ensemble de  $n+1$  points on doit calculer les  $n+1$  polynômes  $\{L_0, L_1, L_2, \dots, L_n\}$ . Si ensuite on ajoute un point d'interpolation, on doit calculer les  $n+2$  polynômes  $\{\tilde{L}_0, \tilde{L}_1, \tilde{L}_2, \dots, \tilde{L}_{n+1}\}$  qui diffèrent tous des  $n+1$  calculés précédemment. La méthode de NEWTON est basée sur le choix d'une autre base de sorte à ce que l'ajout d'un point comporte juste l'ajout d'une fonction de base.

Considérons la famille de polynômes  $\{\omega_0, \omega_1, \omega_2, \dots, \omega_n\}$  où<sup>1</sup>

$$\begin{aligned} \omega_0(x) &= 1, \\ \omega_k(x) &= \prod_{i=0}^{k-1} (x - x_i) = (x - x_{k-1}) \omega_{k-1}(x), \quad \forall k = 1, \dots, n. \end{aligned}$$

Il est facile de vérifier que

- $\omega_k(x) \in \mathbb{R}_n[x]$ ,
- la famille  $\{\omega_0, \omega_1, \omega_2, \dots, \omega_n\}$  est génératrice de  $\mathbb{R}_n[x]$
- la famille  $\{\omega_0, \omega_1, \omega_2, \dots, \omega_n\}$  est libre.

Par conséquent, la famille  $\{\omega_0, \omega_1, \omega_2, \dots, \omega_n\}$  forme une base de  $\mathbb{R}_n[x]$ .

Si on choisit comme base de  $\mathbb{R}_n[x]$  la famille  $\{\omega_0, \omega_1, \omega_2, \dots, \omega_n\}$ , le problème du calcul du polynôme d'interpolation  $p_n$  est alors ramené au calcul des coefficients  $\{\alpha_0, \alpha_1, \alpha_2, \dots, \alpha_n\}$  tels que

$$p_n(x) = \sum_{i=0}^n \alpha_i \omega_i(x).$$

1. Notons que le dernier point  $x_n$  n'intervient pas dans la construction de cette base.

Si on a calculé les  $n+1$  coefficients  $\{\alpha_0, \alpha_1, \alpha_2, \dots, \alpha_n\}$  et on ajoute un point d'interpolation, il n'y a plus à calculer que le coefficient  $\alpha_{n+1}$  car la nouvelle base est déduite de l'autre base en ajoutant simplement le polynôme  $\omega_{n+1}$ .

Pour calculer tous les coefficients on introduit la notion de *différence divisée*: soit  $\{(x_i, y_i)\}_{i=0}^n$  un ensemble de  $n+1$  points distincts.

- La différence divisée d'ordre 1 de  $x_{i-1}$  et  $x_i$  est

$$f[x_{i-1}, x_i] \equiv \frac{y_i - y_{i-1}}{x_i - x_{i-1}}.$$

- La différence divisée d'ordre  $n$  des  $n+1$  points  $x_0, \dots, x_n$  est définie par récurrence en utilisant deux différences divisées d'ordre  $n-1$  comme suit :

$$f[x_0, \dots, x_n] \equiv \frac{f[x_1, \dots, x_n] - f[x_0, \dots, x_{n-1}]}{x_n - x_0}$$

Pour expliciter le processus récursif, les différences divisées peuvent être calculées en les disposant de la manière suivante dans un tableau :

$i$	$x_i$	$y_i$	$f[x_{i-1}, x_i]$	$f[x_{i-2}, x_{i-1}, x_i]$	$f[x_{i-3}, x_{i-2}, x_{i-1}, x_i]$	$f[x_{i-4}, x_{i-3}, x_{i-2}, x_{i-1}, x_i]$	$\dots$
0	$x_0$	$y_0$					
1	$x_1$	$y_1$	$f[x_0, x_1]$				
2	$x_2$	$y_2$	$f[x_1, x_2]$	$f[x_0, x_1, x_2]$			
3	$x_3$	$y_3$	$f[x_2, x_3]$	$f[x_1, x_2, x_3]$	$f[x_0, x_1, x_2, x_3]$		
4	$x_4$	$y_4$	$f[x_3, x_4]$	$f[x_2, x_3, x_4]$	$f[x_1, x_2, x_3, x_4]$	$f[x_0, x_1, x_2, x_3, x_4]$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

Soit  $\{(x_i, y_i)\}_{i=0}^n$  un ensemble de  $n+1$  points distincts. Le polynôme d'interpolation  $p_n$  sous la forme de NEWTON est donné par

$$p_n(x) = \sum_{i=0}^n \omega_i(x) f[x_0, \dots, x_i]$$

où

$$\begin{aligned} \omega_0(x) &= 1, \\ \omega_k(x) &= \prod_{i=1}^{k-1} (x - x_i) = (x - x_{k-1}) \omega_{k-1}(x), \quad \forall k = 1, \dots, n; \\ f[x_k] &= y_k, \quad \forall k = 0, \dots, n, \\ f[x_0, \dots, x_k] &\equiv \frac{f[x_1, \dots, x_k] - f[x_0, \dots, x_{k-1}]}{x_k - x_0}, \quad \forall k = 1, \dots, n. \end{aligned}$$

Comme le montre la définition des différences divisées, des points supplémentaires peuvent être ajoutés pour créer un nouveau polynôme d'interpolation sans recalculer les coefficients. De plus, si un point est modifié, il est inutile de recalculer l'ensemble des coefficients. Autre avantage, si les  $x_i$  sont équirépartis, le calcul des différences divisées devient nettement plus rapide. Par conséquent, l'interpolation polynomiale dans une base de NEWTON est privilégiée par rapport à une interpolation dans la base de LAGRANGE pour des raisons pratiques.

### EXEMPLE

On se propose de calculer le polynôme d'interpolation de l'ensemble de points  $\{(-1, 1), (0, 0), (1, 1)\}$ . On cherche donc  $p_2 \in \mathbb{R}_2[x]$  tel que  $p_2(x_i) = y_i$  pour  $i = 0, \dots, 2$ .

**Méthode directe.** Si on écrit  $p_2(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2$ , on cherche  $\alpha_0, \alpha_1, \alpha_2$  tels que

$$\begin{pmatrix} 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$$

En résolvant ce système linéaire on trouve  $\alpha_0 = 0$ ,  $\alpha_1 = 0$  et  $\alpha_2 = 1$  ainsi  $p_2(x) = x^2$ .

**Méthode de Lagrange.** On a

$$p_2(x) = y_0 L_0(x) + y_1 L_1(x) + y_2 L_2(x) = \frac{x(x-1)}{(-1-0)(-1-1)} + \frac{(x-(-1))(x-0)}{(1-(-1))(1-0)} = \frac{1}{2}x(x-1) + \frac{1}{2}(x+1)x = x^2$$

**Méthode de Newton.** On commence par construire le tableau des différences divisées :

$i$	$x_i$	$y_i$	$f[x_{i-1}, x_i]$	$f[x_{i-2}, x_{i-1}, x_i]$
0	-1	1		
1	0	0	-1	
2	1	1	1	1

On a alors

$$\begin{aligned} p_2(x) &= \sum_{i=0}^2 \omega_i(x) f[x_0, \dots, x_i] \\ &= \omega_0(x) f[x_0] + \omega_1(x) f[x_0, x_1] + \omega_2(x) f[x_0, x_1, x_2] \\ &= \omega_0(x) - \omega_1(x) + \omega_2(x) \\ &= 1 - (x+1) + x(x+1) = x^2. \end{aligned}$$

### EXEMPLE

On se propose de calculer le polynôme d'interpolation de la fonction  $f(x) = \sin(x)$  en les 3 points  $x_i = \frac{\pi}{2}i$  avec  $i = 0, \dots, 2$ . On cherche donc  $p_2 \in \mathbb{R}_2[x]$  tel que  $p_2(x_i) = \sin(x_i)$  pour  $i = 0, \dots, 2$ .

**Méthode directe.** Si on écrit  $p_2(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2$ , on cherche  $\alpha_0, \alpha_1, \alpha_2$  tels que

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & \frac{\pi}{2} & \frac{\pi^2}{4} \\ 1 & \pi & \pi^2 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

En résolvant ce système linéaire<sup>2</sup> on trouve  $\alpha_0 = 0$ ,  $\alpha_1 = \frac{4}{\pi}$  et  $\alpha_2 = -\frac{4}{\pi^2}$  ainsi  $p_2(x) = \frac{4}{\pi}x - \frac{4}{\pi^2}x^2 = \frac{4}{\pi^2}x(\pi - x)$ .

**Méthode de Lagrange.** On a

$$p_2(x) = y_0 L_0(x) + y_1 L_1(x) + y_2 L_2(x) = \frac{x(x-\pi)}{\frac{\pi}{2}(\frac{\pi}{2}-\pi)} = -\frac{4}{\pi^2}x(x-\pi).$$

**Méthode de Newton.** On commence par construire le tableau des différences divisées :

$i$	$x_i$	$y_i$	$f[x_{i-1}, x_i]$	$f[x_{i-2}, x_{i-1}, x_i]$
0	0	0		
1	$\frac{\pi}{2}$	1	$\frac{2}{\pi}$	
2	$\pi$	0	$-\frac{2}{\pi}$	$-\frac{4}{\pi^2}$

On a alors

$$\begin{aligned} p_2(x) &= \sum_{i=0}^2 \omega_i(x) f[x_0, \dots, x_i] \\ &= \omega_0(x) f[x_0] + \omega_1(x) f[x_0, x_1] + \omega_2(x) f[x_0, x_1, x_2] \\ &= \frac{2}{\pi}\omega_1(x) - \frac{4}{\pi^2}\omega_2(x) \\ &= \frac{2}{\pi}x - \frac{4}{\pi^2}x\left(x - \frac{\pi}{2}\right) \end{aligned}$$

2. Par la méthode du pivot de Gauss on obtient

$$\left( \begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ 1 & \frac{\pi}{2} & \frac{\pi^2}{4} & 1 \\ 1 & \pi & \pi^2 & 0 \end{array} \right) \xrightarrow{L_2 \leftarrow L_2 - L_1} \left( \begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ 0 & \frac{\pi}{2} & \frac{\pi^2}{4} & 1 \\ 1 & \pi & \pi^2 & 0 \end{array} \right) \xrightarrow{L_3 \leftarrow L_3 - 2L_2} \left( \begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ 0 & \frac{\pi}{2} & \frac{\pi^2}{4} & 1 \\ 0 & 0 & \frac{\pi^2}{2} & -2 \end{array} \right)$$

$$= -\frac{4}{\pi^2} x(x - \pi).$$

Maintenant on veut calculer le polynôme d'interpolation de la même fonction en les 4 points  $x_i = \frac{\pi}{2} i$  avec  $i = 0, \dots, 3$ , i.e. on a juste ajouté le point  $x = 3\pi/2$ . On cherche donc  $p_3 \in \mathbb{R}_3[x]$  tel que  $p_3(x_i) = \sin(x_i)$  pour  $i = 0, \dots, 3$ .

**Méthode directe.** Si on écrit  $p_3(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3$ , on cherche  $\alpha_0, \alpha_1, \alpha_2, \alpha_3$  tels que

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & \frac{\pi}{2} & \frac{\pi^2}{4} & \frac{\pi^3}{8} \\ 1 & \pi & \pi^2 & \pi^3 \\ 1 & \frac{3\pi}{2} & \frac{9\pi^2}{4} & \frac{27\pi^3}{8} \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ -1 \end{pmatrix}$$

En résolvant ce système linéaire on trouve  $\alpha_0 = 0$ ,  $\alpha_1 = \frac{16}{3\pi}$ ,  $\alpha_2 = -\frac{8}{\pi^2}$  et  $\alpha_3 = \frac{8}{3\pi^3}$ .

**Méthode de Lagrange.** On a

$$\begin{aligned} p_3(x) &= y_0 L_0(x) + y_1 L_1(x) + y_2 L_2(x) + y_3 L_3(x) = \frac{x(x-\pi)(x-\frac{3\pi}{2})}{\frac{\pi}{2}(\frac{\pi}{2}-\pi)(\frac{\pi}{2}-\frac{3\pi}{2})} - \frac{x(x-\frac{\pi}{2})(x-\pi)}{\frac{3\pi}{2}(\frac{3\pi}{2}-\frac{\pi}{2})(\frac{3\pi}{2}-\pi)} \\ &= \frac{4}{\pi^3} x(x-\pi)\left(x-\frac{3\pi}{2}\right) - \frac{4}{3\pi^3} x\left(x-\frac{\pi}{2}\right)(x-\pi). \end{aligned}$$

**Méthode de Newton.** Il suffit de calculer une différence divisée en plus, i.e. ajouter une ligne au tableau :

$i$	$x_i$	$y_i$	$f[x_{i-1}, x_i]$	$f[x_{i-2}, x_{i-1}, x_i]$	$f[x_{i-3}, x_{i-2}, x_{i-1}, x_i]$
0	0	0			
1	$\frac{\pi}{2}$	1		$\frac{2}{\pi}$	
2	$\pi$	0	$-\frac{2}{\pi}$	$-\frac{4}{\pi^2}$	
3	$\frac{3\pi}{2}$	-1	$-\frac{2}{\pi}$	0	$\frac{8}{3\pi^3}$

On a alors

$$\begin{aligned} p_3(x) &= \sum_{i=0}^3 \omega_i(x) f[x_0, \dots, x_i] \\ &= p_2(x) + \omega_3(x) f[x_0, x_1, x_2, x_3] \\ &= -\frac{4}{\pi^2} x(x-\pi) + \frac{8}{3\pi^3} \omega_3(x) \\ &= -\frac{4}{\pi^2} x(x-\pi) + \frac{8}{3\pi^3} x\left(x-\frac{\pi}{2}\right)(x-\pi) \\ &= \frac{8}{3\pi^3} x(x^2 - 3\pi x + 2\pi^2). \end{aligned}$$

### Remarque

Si  $n$  est petit il est souvent plus simple de calculer directement les coefficients  $a_0, a_1, \dots, a_n$  en résolvant le système linéaire (2.2).

## 2.2 Splines : interpolation composite

On a mis en évidence le fait qu'on ne peut pas garantir la convergence uniforme du polynôme interpolatoire de LAGRANGE vers  $f$  quand les nœuds d'interpolation sont équirépartis. L'interpolation de LAGRANGE de bas degré est cependant suffisamment précise quand elle est utilisée sur des intervalles assez petits, y compris avec des nœuds équirépartis (ce qui est commode en pratique). Il est donc naturel d'introduire une partition de  $[a; b]$  en  $n$  sous-intervalles  $[x_i, x_{i+1}]$ , tels que  $[a; b] = \cup_{0 \leq i \leq n-1} [x_i, x_{i+1}]$  et d'utiliser l'interpolation de LAGRANGE sur chaque sous-intervalle  $[x_i, x_{i+1}]$  en utilisant  $m$  nœuds équirépartis avec  $m$  petit (généralement  $m = 1$  ou 3).

Étant donné une distribution (non nécessairement uniforme) de nœuds  $x_0 < x_1 < \dots < x_n$ , on approche  $f$  par une fonction continue qui, sur chaque intervalle  $[x_i, x_{i+1}]$ , est définie par le segment joignant les deux points  $(x_i, f(x_i))$  et  $(x_{i+1}, f(x_{i+1}))$ . Cette fonction est appelée interpolation linéaire par morceaux (ou *spline* linéaire).

### Définition 2.1 (Splines linéaires)

Étant donné  $n+1$  points distincts  $x_0, \dots, x_n$  de  $[a; b]$  avec  $a = x_0 < x_1 < \dots < x_n = b$ , la fonction  $\ell: [a; b] \rightarrow \mathbb{R}$  est une spline linéaire relative aux nœuds  $\{x_i\}$  si

$$\begin{cases} \ell(x)|_{[x_i; x_{i+1}]} \in \mathbb{R}_1, & i = 1, 1, \dots, n-1, \\ \ell \in \mathcal{C}^0([a; b]). \end{cases}$$

Autrement dit, dans chaque sous-intervalle  $[x_i; x_{i+1}]$ , la fonction  $\ell: [x_i, x_{i+1}] \rightarrow \mathbb{R}$  est le segment qui connecte le point  $(x_i, y_i)$  au point  $(x_{i+1}, y_{i+1})$ ; elle s'écrit donc

$$\ell(x)|_{[x_i; x_{i+1}]} = y_i + \frac{y_{i+1} - y_i}{x_{i+1} - x_i}(x - x_i)$$

Il est intéressant de noter que la commande `plot(x, y)`, utilisée pour afficher le graphe d'une fonction  $f$  sur un intervalle donné  $[a, b]$ , remplace en fait la fonction par une interpolée linéaire par morceaux, les points d'interpolation étant les composantes du vecteur `x`.

Le principale défaut de cette interpolation par morceaux est que  $\ell$  n'est que continue. Or, dans des nombreuses applications, il est préférable d'utiliser des fonctions ayant au moins une dérivée continue. On peut construire pour cela une fonction  $s_3$  comme l'interpolation d'HERMITE des points  $(x_i, f(x_i), f'(x_i))$  et  $(x_{i+1}, f(x_{i+1}), f'(x_{i+1}))$  sur chaque  $[x_i; x_{i+1}]$  pour  $i = 1, 1, \dots, n-1$ .

TO DO interpolation d'Hermite, splines cubiques

## 2.3 Interpolation Trigonométrique

On veut approcher une fonction périodique  $f: [0; 2\pi] \rightarrow \mathbb{C}$ , i.e. satisfaisant  $f(0) = f(2\pi)$ , par un polynôme trigonométrique  $\tilde{f}$ , i.e. une combinaison linéaire de sinus et de cosinus, qui interpole  $f$  aux  $n+1$  nœuds équidistants  $x_j = jh \in [0; 2\pi]$  avec  $j = 0, \dots, n$  et  $h = \frac{2\pi}{n+1}$ . On remarque que le point  $2\pi$  est omis car redondant avec le point  $x=0$  étant donné que  $f(0) = f(2\pi)$ .

La fonction d'interpolation trigonométrique  $\tilde{f}$  peut s'écrire comme

$$\tilde{f}(x) = a_0 + \sum_{k=1}^K a_k \cos(kx) + b_k \sin(kx)$$

dont les inconnues sont le coefficient complexes  $a_0$  et les  $2K$  coefficients  $a_k$  et  $b_k$ . On peut remarquer que  $\tilde{f}$  s'apparente à une série de FOURIER tronquée, i.e. au lieu de sommer jusqu'à l'infini on tronque la somme à l'entier  $K$ .

Rappels :

$$\begin{cases} \cos(kx) = \frac{e^{ikx} + e^{-ikx}}{2}, \\ \sin(kx) = \frac{e^{ikx} - e^{-ikx}}{2i} = -i \frac{e^{ikx} - e^{-ikx}}{2}, \end{cases} \quad \text{et} \quad \begin{cases} e^{ikx} = \cos(kx) + i \sin(kx), \\ e^{-ikx} = \cos(kx) - i \sin(kx). \end{cases}$$

Ainsi

$$\begin{aligned} \tilde{f}(x) &= a_0 + \sum_{k=1}^K a_k \frac{e^{ikx} + e^{-ikx}}{2} - i b_k \frac{e^{ikx} - e^{-ikx}}{2} \\ &= a_0 + \sum_{k=1}^K \underbrace{\frac{a_k - i b_k}{2}}_{c_k} e^{ikx} + \underbrace{\frac{a_k + i b_k}{2}}_{c_{-k}} e^{i(-k)x} = \sum_{k=-K}^K c_k e^{ikx} \end{aligned}$$

les inconnues sont maintenant les  $2K+1$  coefficients  $c_k \in \mathbb{C}$  et l'on a les relations

$$\begin{cases} c_0 = a_0, \\ c_k = \frac{a_k - i b_k}{2}, & k = 1, \dots, K \\ c_{-k} = \overline{c_k} = \frac{a_k + i b_k}{2}, & k = 1, \dots, K \end{cases} \iff \begin{cases} a_0 = c_0, \\ a_k = c_k + c_{-k}, & k = 1, \dots, K \\ b_k = i(c_k - c_{-k}), & k = 1, \dots, K. \end{cases}$$

Une autre écriture souvent utilisée se base sur l'écriture exponentielle des coefficients  $c_k$ : pour tout  $k$ ,  $c_k \in \mathbb{C}$  peut s'écrire comme  $c_k = \frac{1}{2} r_k e^{i\varphi_k}$  ainsi  $c_{-k} = \overline{c_k} = \frac{1}{2} r_k e^{-i\varphi_k}$  et on trouve

$$a_k \cos(kx) + b_k \sin(kx) = c_k e^{ikx} + \overline{c_k} e^{-ikx} = \frac{1}{2} r_k e^{i\varphi_k} e^{ikx} + \frac{1}{2} r_k e^{-i\varphi_k} e^{-ikx} = \frac{1}{2} r_k \left( e^{i(\varphi_k+kx)} + e^{-i(\varphi_k+kx)} \right) = r_k \cos(kx + \varphi_k).$$

Ainsi les inconnues sont maintenant le coefficient  $a_0$  et les  $2K$  couples “amplitude, phase”  $(r_k, \varphi_k) \in \mathbb{R}$ :

$$\tilde{f}(x) = a_0 + \sum_{k=1}^K r_k \cos(kx + \varphi_k)$$

En écrivant les  $n+1$  conditions d’interpolation aux nœuds  $x_j$  on trouve

$$f(x_j) = \tilde{f}(x_j) = \sum_{k=-K}^K c_k e^{ikx_j}.$$

Quand  $n$  est pair, on pose  $K = n/2$  ainsi nous avons  $n+1$  conditions d’interpolation et  $2K+1 = n+1$  inconnues; quand  $n$  est impair, on pose  $K = (n+1)/2$  ainsi nous avons  $n+1$  conditions d’interpolation et  $2K+1 = n+2$  inconnues, pour fermer le système on ajoute alors la condition  $c_K = 0$ . Pour uniformiser la notation dans ces deux cas, nous pouvons écrire  $M = n/2$  et

$$\tilde{f}(x) = \sum_{k=-(M+\mu)}^M c_k e^{ikx}, \quad \mu = \begin{cases} 0 & \text{si } n \text{ est pair,} \\ 1 & \text{si } n \text{ est impair,} \end{cases}$$

et les  $n+1$  conditions d’interpolation aux nœuds  $x_j = jh$  donnent les  $n+1$  conditions

$$f(x_j) = \tilde{f}(x_j) = \sum_{k=-(M+\mu)}^M c_k e^{ikjh}.$$

Pour calculer les  $n+1$  inconnues  $\{c_k\}_{k=-M-\mu}^M$ , on multiplie cette équation par  $e^{-imjh}$  où  $m = -M-\mu, \dots, M$  et on somme sur  $j$ :

$$\sum_{j=0}^n (f(x_j) e^{-imjh}) = \sum_{j=0}^n \left( \sum_{k=-M-\mu}^M c_k e^{i(k-m)jh} \right).$$

En échangeant l’ordre de sommation on obtient

$$\sum_{j=0}^n (f(x_j) e^{-imjh}) = \sum_{k=-M-\mu}^M \left( c_k \left( \sum_{j=0}^n e^{i(k-m)jh} \right) \right).$$

On se rappelle que  $\sum_{j=0}^n q^j = (n+1)$  si  $q = 1$  et  $\sum_{j=0}^n q^j = \frac{1-q^{n+1}}{1-q}$  si  $q \neq 1$ , ainsi en prenant  $q = e^{i(k-m)h}$  on a

$$\sum_{j=0}^n (e^{i(k-m)h})^j = (n+1)\delta_{km}$$

car  $\sum_{j=0}^n (e^{i(k-m)h})^j = n+1$  si  $k = m$  et si  $k \neq m$  alors

$$\sum_{j=0}^n (e^{i(k-m)h})^j = \frac{1 - (e^{i(k-m)h})^{n+1}}{1 - (e^{i(k-m)h})} = \frac{1 - e^{i(k-m)(n+1)h}}{1 - (e^{i(k-m)h})} = \frac{1 - e^{i(k-m)2\pi}}{1 - (e^{i(k-m)h})} = \frac{1 - \cos((k-m)2\pi) - i \sin((k-m)2\pi)}{1 - (e^{i(k-m)h})} = 0.$$

Donc

$$\sum_{j=0}^n (f(x_j) e^{-imjh}) = (n+1) \sum_{k=-M-\mu}^M \delta_{km} c_k$$

i.e. seul le terme  $k = m$  est à prendre en considération

$$\sum_{j=0}^n (f(x_j) e^{-imjh}) = (n+1)c_m \quad m = -M-\mu, \dots, M.$$

Soit  $\{(x_j = jh, f(x_j))\}_{j=0}^n$  un ensemble de  $n+1$  points avec  $h = 2\pi/(n+1)$  et  $f: [0; 2\pi] \rightarrow \mathbb{C}$  une fonction périodique. Le polynôme trigonométrique d’interpolation  $\tilde{f}$  est donné par

$$\tilde{f}(x) = \sum_{k=-M-\mu}^M c_k e^{-ikx}, \quad (M, \mu) = \begin{cases} (n/2, 0) & \text{si } n \text{ est pair,} \\ ((n-1)/2, 1) & \text{si } n \text{ est impair,} \end{cases}$$

et, pour  $k = -(M + \mu) \dots M$ ,

$$c_k = \frac{1}{n+1} \sum_{j=0}^n f(x_j) e^{ikx_j}.$$

De manière équivalente on peut écrire

$$\tilde{f}(x) = a_0 + \sum_{k=1}^{M+\mu} a_k \cos(kx) + b_k \sin(kx), \quad (M, \mu) = \begin{cases} (n/2, 0) & \text{si } n \text{ est pair,} \\ ((n-1)/2, 1) & \text{si } n \text{ est impair,} \end{cases}$$

avec

$$\begin{cases} a_0 = \frac{1}{n+1} \sum_{j=0}^n f(x_j) \\ a_k = \frac{2}{n+1} \sum_{j=0}^n f(x_j) \cos(kx_j), \quad k = 1, \dots, M + \mu, \\ b_k = \frac{2}{n+1} \sum_{j=0}^n f(x_j) \sin(kx_j), \quad k = 1, \dots, M + \mu, \\ a_{M+\mu} = i b_{M+\mu} \end{cases} \quad \text{si } \mu = 1.$$

Il est intéressant de noter que l'expression de  $c_k$  est une approximation de l'intégrale  $\frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-ikx} dx$  par la méthode des rectangles à gauche composite. De la même manière, les coefficients  $a_k$  et  $b_k$  sont des approximations des intégrales  $\frac{1}{\pi} \int_0^{2\pi} f(x) \cos(kx) dx$  et  $\frac{1}{\pi} \int_0^{2\pi} f(x) \sin(kx) dx$  respectivement. Vu que ces intégrales définissent précisément les coefficients de FOURIER, on déduit que nos sommes sont des approximations des coefficients de FOURIER et on parle alors d'une transformation de FOURIER discrète. Le calcul des coefficients  $c_k$  peut ainsi être effectué en utilisant la transformation de Fourier rapide (FFT).

Notons que si  $f$  est une fonction à valeurs réelles, alors  $c_{-k} = \overline{c_k}$  et donc  $\tilde{f}$  aussi est une fonction à valeurs réelles.

### EXEMPLE

Considérons la fonction  $f: [0; 2\pi] \rightarrow \mathbb{R}$  définie par  $f(x) = x(x - 2\pi)e^{-x}$ . On a bien  $f(0) = f(2\pi)$ .

- \* On se propose de calculer  $\tilde{f}(x)$  lorsque  $n = 1$ . On a  $x_j = jh$  avec  $j = 0, 1$  et  $h = \pi$ . On interpole alors les deux points  $\{(0, f(0)), (\pi, f(\pi))\} = \{(0, 0), (\pi, -\pi^2 e^{-\pi})\}$ .  
 $n$  étant impair,  $M = (n-1)/2 = 0$  et  $\mu = 1$  et

$$\tilde{f}(x) = \sum_{k=-1}^0 c_k e^{ikx} = c_{-1} e^{-ix} + c_0$$

On doit alors calculer les deux coefficients de FOURIER  $c_{-1}$  et  $c_0$  :

$$\begin{aligned} c_{-1} &= \frac{1}{n+1} \sum_{j=0}^1 f(x_j) e^{-ix_j} = \frac{1}{2} (f(x_0) e^{-ix_0} + f(x_1) e^{-ix_1}) = \frac{1}{2} (-\pi^2 e^{-\pi} e^{-i\pi}) = \frac{\pi^2 e^{-\pi}}{2} \\ c_0 &= \frac{1}{n+1} \sum_{j=0}^1 f(x_j) = \frac{1}{2} (f(x_0) + f(x_1)) = \frac{1}{2} (-\pi^2 e^{-\pi}) = -\frac{\pi^2 e^{-\pi}}{2} \end{aligned}$$

ainsi

$$\tilde{f}(x) = \frac{\pi^2 e^{-\pi}}{2} (e^{-ix} - 1).$$

- \* On se propose de calculer  $\tilde{f}(x)$  lorsque  $n = 2$ . On a  $x_j = jh$  avec  $j = 0, 1, 2$  et  $h = \frac{2\pi}{3}$ . On interpole alors les trois points  $\{(0, f(0)), (\frac{2\pi}{3}, f(\frac{2\pi}{3})), (\frac{4\pi}{3}, f(\frac{4\pi}{3}))\} = \{(0, 0), (\frac{2\pi}{3}, -\frac{8\pi^2}{9} e^{-\frac{2\pi}{3}}), (\frac{4\pi}{3}, -\frac{8\pi^2}{9} e^{-\frac{4\pi}{3}})\}$ .  
 $n$  étant pair,  $M = n/2 = 1$  et  $\mu = 0$  et

$$\tilde{f}(x) = \sum_{k=-1}^1 c_k e^{ikx} = c_{-1} e^{-ix} + c_0 + c_1 e^{ix}$$

On doit alors calculer les trois coefficients de FOURIER  $c_{-1}$ ,  $c_0$  et  $c_1$  :

$$\begin{aligned} c_{-1} &= \frac{1}{n+1} \sum_{j=0}^2 f(x_j) e^{-ix_j} = \frac{1}{3} (f(x_0) e^{-ix_0} + f(x_1) e^{-ix_1} + f(x_2) e^{-ix_2}) = \frac{1}{3} \left( -\frac{8\pi^2}{9} e^{-(i+1)\frac{2\pi}{3}} - \frac{8\pi^2}{9} e^{-(i+1)\frac{4\pi}{3}} \right) \\ &= -\frac{8\pi^2}{27} \left( e^{-(i+1)\frac{2\pi}{3}} + e^{-(i+1)\frac{4\pi}{3}} \right) = -\frac{8\pi^2}{27} e^{-(i+1)\frac{2\pi}{3}} \left( 1 + e^{-(i+1)\frac{2\pi}{3}} \right) \\ c_0 &= \frac{1}{n+1} \sum_{j=0}^2 f(x_j) = \frac{1}{3} (f(x_0) + f(x_1) + f(x_2)) = \frac{1}{3} \left( -\frac{8\pi^2}{9} e^{-\frac{2\pi}{3}} - \frac{8\pi^2}{9} e^{-\frac{4\pi}{3}} \right) = -\frac{8\pi^2}{27} e^{-\frac{2\pi}{3}} \left( 1 + e^{-\frac{2\pi}{3}} \right) \end{aligned}$$

$$\begin{aligned}
c_1 &= \frac{1}{n+1} \sum_{j=0}^2 f(x_j) e^{ix_j} = \frac{1}{3} \left( f(x_0) e^{ix_0} + f(x_1) e^{ix_1} + f(x_2) e^{ix_2} \right) = \frac{1}{3} \left( -\frac{8\pi^2}{9} e^{(i-1)\frac{2\pi}{3}} - \frac{8\pi^2}{9} e^{(i-1)\frac{4\pi}{3}} \right) \\
&= -\frac{8\pi^2}{27} \left( e^{(i-1)\frac{2\pi}{3}} + e^{(i-1)\frac{4\pi}{3}} \right) = -\frac{8\pi^2}{27} e^{(i-1)\frac{2\pi}{3}} \left( 1 + e^{(i-1)\frac{2\pi}{3}} \right)
\end{aligned}$$

ainsi

$$\begin{aligned}
\tilde{f}(x) &= -\frac{8\pi^2}{27} e^{-(i+1)\frac{2\pi}{3}} \left( 1 + e^{-(i+1)\frac{2\pi}{3}} \right) e^{-ix} - \frac{8\pi^2}{27} e^{-\frac{2\pi}{3}} \left( 1 + e^{-\frac{2\pi}{3}} \right) - \frac{8\pi^2}{27} e^{(i-1)\frac{2\pi}{3}} \left( 1 + e^{(i-1)\frac{2\pi}{3}} \right) e^{ix} \\
&= -\frac{8\pi^2}{27} e^{-\frac{2\pi}{3}} \left[ e^{-i\frac{2\pi}{3}} \left( 1 + e^{-(i+1)\frac{2\pi}{3}} \right) e^{-ix} + \left( 1 + e^{-\frac{2\pi}{3}} \right) + e^{i\frac{2\pi}{3}} \left( 1 + e^{(i-1)\frac{2\pi}{3}} \right) e^{ix} \right] \\
&= -\frac{8\pi^2}{27} e^{-\frac{2\pi}{3}} \left[ e^{-i(\frac{2\pi}{3}+x)} \left( 1 + e^{-(i+1)\frac{2\pi}{3}} \right) + \left( 1 + e^{-\frac{2\pi}{3}} \right) + e^{i(\frac{2\pi}{3}+x)} \left( 1 + e^{(i-1)\frac{2\pi}{3}} \right) \right].
\end{aligned}$$

TO DO : Passer à une fonction périodique sur un autre intervalle. Expliquer la transformation de Fourier, la transformation de Fourier discrète (= série?) et l'algorithme de transformation de Fourier rapide (FFT) et les liens avec l'interpolation. Passer à une fonction non périodique et/ou non continue. Illustrer la convergence uniforme en opposition à celle non uniforme de l'interpolation polynomiale.





## Exercices



### Interpolation polynomiale

#### ★ Exercice 2.1

On se propose d'écrire trois **function** pour évaluer le polynôme d'interpolation d'un ensemble de points, une pour chaque méthode vue en cours (base canonique, base de LAGRANGE et base de NEWTON). Chaque **function** prend en entrée P une matrice de  $n$  lignes et 2 colonnes qui contient les points d'interpolation et x le vecteur contenant les points où on veut évaluer le polynôme d'interpolation et elle donne en sortie y le vecteur contenant l'évaluation du polynôme d'interpolation.

#### Correction

- ① Dans le fichier `naive.m` on définit la fonction suivante

```
function [y]=naive(P,x)
[1,c]=size(P);
V = ones(1,1);
V(:,2:1) = P(:,1).^(1:c-1);
alpha = V\P(:,2);
y=zeros(size(x));
for i=1:c
    y+=alpha(i)*x.^(i-1);
end
end
```

puis on teste comme suit : le seul polynôme de degré au plus 2 qui interpole l'ensemble de points  $\{(-2, 4), (0, 0), (1, 1)\}$  est la parabole d'équation  $p(x) = x^2$  et lorsqu'on évalue  $p$  en  $-1$ , en  $0$  et en  $2$  on trouve respectivement  $1$ ,  $0$  et  $4$  :

```
>> P=[-2 4; 0 0; 1 1];
>> y=naive(P,[-1 0 2])
y =
1 0 4
```

Remarque : la commande  $V(:,2:1) = P(:,1).^(1:c-1)$  correspond à la boucle

```
for j=1:c
    V(:,j) = P(:,1).^(j-1);
end
```

Profitons de cet exercice pour décrire une méthode pour évaluer efficacement la valeur d'un polynôme en un point donné  $x$ . D'un point de vue algébrique, nous pouvons écrire

$$p(x) = \sum_{i=1}^n \alpha_i x^{i-1} = \alpha_1 + x \left( \alpha_2 + x \left( \alpha_3 + \cdots + x (\alpha_{n-1} + \alpha_n x) \dots \right) \right).$$

Tandis que  $\sum_{i=1}^n \alpha_i x^{i-1}$  nécessite  $n$  sommes et  $2n - 1$  produits pour évaluer le polynôme (pour un  $x$  donné), la deuxième écriture ne requiert que  $n$  sommes et  $n$  produits. Cette dernière expression, parfois appelée méthode des produits imbriqués, est la base de l'algorithme de HÖRNER. Celui-ci permet d'évaluer de manière efficace un polynôme en un point en utilisant l'algorithme de division synthétique suivant :

$$\begin{cases} b_n = \alpha_n \\ b_k = \alpha_k + x b_{k+1} \quad \text{pour } k = n-1, \dots, 1 \end{cases}$$

et  $b_0 = p(x)$ . On modifie alors notre fonction comme suit

```
function [y]=naive(P,x)
[1,c]=size(P);
V = ones(1,1);
V(:,2:1) = P(:,1).^(1:c-1);
alpha = V\P(:,2);
y=zeros(size(x));
for k=1:-1:1
    y=alpha(k)+x.*y;
end
end
```

On peut décomposer notre fonction en deux fonctions : la première rend les coefficients du polynôme dans la base canonique, la deuxième évalue le polynôme lorsqu'on connaît ces coefficients :

```

function [alpha]=naivePoly(P)
    [l,c]=size(P);
    V = ones(1,1);
    V(:,2:l) = P(:,1).^(1:l-1);
    alpha = V\P(:,2);
end

function [y]=naiveEval(alpha,x)
    y=zeros(size(x));
    for k=size(alpha):-1:1
        y=alpha(k)+x.*y;
    end
end

```

- ② Dans le fichier `lagrange.m` on définit la fonction suivante

```

function [y]=lagrange(P,x)
    [l,c]=size(P);
    y=zeros(size(x));
    for i=1:l
        Li=ones(size(x));
        for j=[1:i-1, i+1:l] % pour eviter le test "if (j~=i)"
            Li.*=(x-P(j,1))/(P(i,1)-P(j,1));
        end
        y+=P(i,2)*Li;
    end
end

```

puis on teste comme au point précédent :

```

>> P=[-2 4; 0 0; 1 1];
>> y=lagrange(P, [-1 0 2])
y =

```

1 0 4

- ③ Dans le fichier `newton.m` on définit la fonction suivante

```

function [y]=newton(P,x)
    [l,c]=size(P);
    % calcul des coefficients beta(i)=A(i,i)
    A=zeros(l);
    A(:,1)=P(:,2);
    for j=2:l
        A(j:1,j)=(A(j:1,j-1)-A(j-1:l-1,j-1))./(P(j:1,1)-P(1:l-j+1,1));
    end
    % evaluation des polynomes omega(i) en x et calcul de p(x)
    omegai=ones(size(x));
    y=A(1,1)*omegai;
    for i=2:l
        omegai=omegai.*(x-P(i-1,1));
        y=y+A(i,i)*omegai;
    end
end

```

puis on teste comme au point précédent :

```

>> P=[-2 4; 0 0; 1 1];
>> y=newton(P, [-1 0 2])
y =

```

1 0 4

Remarque : la commande  $A(j:1,j)=(A(j:1,j-1)-A(j-1:l-1,j-1))./(P(j:1,1)-P(1:l-j+1,1))$ ; correspond à la boucle

```

for i=j:1
    A(i,j)=(A(i,j-1)-A(i-1,j-1))/(P(i,1)-P(i-j+1,1));
end

```

D'un point de vue algébrique, nous pouvons écrire

$$p(x) = \sum_{i=1}^n \beta_i \omega_i(x) = \beta_1 + \beta_2(x - x_1) + \beta_3(x - x_1)(x - x_2) + \cdots + \beta_n(x - x_1) \cdots (x - x_{n-1})$$

$$= \beta_1 + (x - x_1) \left( \beta_2 + (x - x_2) \left( \beta_3 + \cdots + (x - x_{n-2}) (\beta_{n-1} + \beta_n (x - x_{n-1})) \dots \right) \right).$$

Celui-ci permet d'évaluer de manière efficace un polynôme en un point en utilisant l'algorithme de division synthétique suivant :

$$\begin{cases} b_n = (x - x_{n-1})\beta_n \\ b_k = \beta_k + (x - x_k)b_{k+1} \text{ pour } k = n-1, \dots, 1 \end{cases}$$

et  $b_0 = p(x)$ . On modifie alors notre fonction comme suit

```
function [y]=newton(P,x)
[l,c]=size(P);
% calcul des coefficients beta(i)=A(i,i)
A=zeros(1);
A(:,1)=P(:,2);
for j=2:1
    A(j:1,j)=(A(j:1,j-1)-A(j-1:1-1,j-1))./(P(j:1,1)-P(1:1-j+1,1));
end
% calcul de p(x)
y=zeros(size(x));
for k=l:-1:1
    y=A(k,k)+(x-P(k,1)).*y;
end
end
```

## Exercice 2.2

Construire le polynôme  $P$  qui interpole les points  $(0, 2), (1, 1), (2, 2)$  et  $(3, 3)$ .

### Correction

On cherche un polynôme de degré au plus 3 tel que  $P(0) = 2$ ,  $P(1) = 1$ ,  $P(2) = 2$  et  $P(3) = 3$ . Construire  $P$  signifie trouver ses coordonnées dans une base de  $\mathbb{R}_3[x]$ . On considère quatre méthodes qui sont basées sur trois choix différents de bases de  $\mathbb{R}_3[x]$  :

- **Méthode astucieuse**

On remarque que les points  $(1, 1), (2, 2)$  et  $(3, 3)$  sont alignés, ainsi le polynôme  $Q(x) = x$  de  $\mathbb{R}_2[x]$  interpole ces points.

Introduisons le polynôme  $D(x) = P(x) - Q(x)$  de  $\mathbb{R}_3[x]$ . Par construction, ce polynôme s'annule en  $x = 1, x = 2$  et  $x = 3$ , donc  $D(x) = \lambda(x - 1)(x - 2)(x - 3)$ . De plus,  $D(0) = P(0) - Q(0) = 2$  mais aussi  $D(0) = -6\lambda$  donc  $\lambda = -1/3$  et on conclut que

$$P(x) = D(x) + Q(x) = -\frac{1}{3}(x - 1)(x - 2)(x - 3) + x.$$

- **Méthode directe (naïve)**

On considère  $\mathcal{C} = \{1, x, x^2, x^3\}$  la base canonique de  $\mathbb{R}_3[x]$  et on cherche  $(a_0, a_1, a_2, a_3) = \text{coord}(P, \mathcal{C})$ , i.e.  $a_0, a_1, a_2, a_3$  tels que  $P(x) = \sum_{i=0}^3 a_i x^i$ .

Il s'agit de trouver les 4 coefficients  $a_0, a_1, a_2$  et  $a_3$  solution du système linéaire

$$\begin{cases} P(0) = 2 \\ P(1) = 1 \\ P(2) = 2 \\ P(3) = 3 \end{cases} \iff \begin{cases} a_0 + a_1 \cdot 0 + a_2 \cdot 0^2 + a_3 \cdot 0^3 = 2 \\ a_0 + a_1 \cdot 1 + a_2 \cdot 1^2 + a_3 \cdot 1^3 = 1 \\ a_0 + a_1 \cdot 2 + a_2 \cdot 2^2 + a_3 \cdot 2^3 = 2 \\ a_0 + a_1 \cdot 3 + a_2 \cdot 3^2 + a_3 \cdot 3^3 = 3 \end{cases} \iff \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \\ 1 & 3 & 9 & 27 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 2 \\ 3 \end{pmatrix}$$

On peut utiliser la méthode de GAUSS-JORDAN :

$$\begin{array}{c|cccc|c} 1 & 0 & 0 & 0 & 2 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 & 2 \\ 1 & 3 & 9 & 27 & 3 \end{array} \xrightarrow{\begin{array}{l} L_2 \leftarrow L_2 - L_1 \\ L_3 \leftarrow L_3 - L_1 \\ L_4 \leftarrow L_4 - L_1 \end{array}} \begin{array}{c|cccc|c} 1 & 0 & 0 & 0 & 2 \\ 0 & 1 & 1 & 1 & -1 \\ 0 & 2 & 4 & 8 & 0 \\ 0 & 3 & 9 & 27 & 1 \end{array} \xrightarrow{\begin{array}{l} L_1 \leftarrow L_1 - 0L_2 \\ L_3 \leftarrow L_3 - 2L_2 \\ L_4 \leftarrow L_4 - 3L_2 \end{array}} \begin{array}{c|cccc|c} 1 & 0 & 0 & 0 & 2 \\ 0 & 1 & 1 & 1 & -1 \\ 0 & 0 & 2 & 6 & 2 \\ 0 & 0 & 6 & 24 & 4 \end{array} \\ \xrightarrow{\begin{array}{l} L_1 \leftarrow L_1 - 0L_3 \\ L_2 \leftarrow L_2 - L_3/2 \\ L_4 \leftarrow L_4 - 3L_3 \end{array}} \begin{array}{c|cccc|c} 1 & 0 & 0 & 0 & 2 \\ 0 & 1 & 0 & -2 & -2 \\ 0 & 0 & 2 & 6 & 2 \\ 0 & 0 & 0 & 6 & -2 \end{array} \xrightarrow{\begin{array}{l} L_1 \leftarrow L_1 - 0L_4 \\ L_2 \leftarrow L_2 + L_4/3 \\ L_3 \leftarrow L_3 - L_4 \end{array}} \begin{array}{c|cccc|c} 1 & 0 & 0 & 0 & 2 \\ 0 & 1 & 0 & 0 & -8/3 \\ 0 & 0 & 2 & 0 & 4 \\ 0 & 0 & 0 & 6 & -2 \end{array} \end{array}$$

donc  $a_3 = -\frac{1}{3}$ ,  $a_2 = 2$ ,  $a_1 = -\frac{8}{3}$  et  $a_0 = 2$  et on trouve  $P(x) = 2 - \frac{8}{3}x + 2x^2 - \frac{1}{3}x^3$ :

```
P=[0 2; 1 1; 2 2; 3 3];
alpha=naivePoly(P)
x=[0:0.1:3];
y=naiveEval(alpha,x);
plot(x,y)
```

*Remarque :* dans ce cas particulier, le système s'écrit

$$\begin{cases} a_0 = 2 \\ a_0 + a_1 \cdot 1 + a_2 \cdot 1^2 + a_3 \cdot 1^3 = 1 \\ a_0 + a_1 \cdot 2 + a_2 \cdot 2^2 + a_3 \cdot 2^3 = 2 \\ a_0 + a_1 \cdot 3 + a_2 \cdot 3^2 + a_3 \cdot 3^3 = 3 \end{cases}$$

ainsi on peut déjà poser  $a_0 = 2$  et résoudre le système linéaire réduit suivant :

$$\begin{cases} a_1 \cdot 1 + a_2 \cdot 1^2 + a_3 \cdot 1^3 = -1 \\ a_1 \cdot 2 + a_2 \cdot 2^2 + a_3 \cdot 2^3 = 0 \\ a_1 \cdot 3 + a_2 \cdot 3^2 + a_3 \cdot 3^3 = 1 \end{cases}$$

On peut utiliser la méthode de GAUSS-JORDAN :

$$\left( \begin{array}{ccc|c} 1 & 1 & 1 & -1 \\ 2 & 4 & 8 & 0 \\ 3 & 9 & 27 & 1 \end{array} \right) \xrightarrow{\substack{L_2 \leftarrow L_2 - 2L_1 \\ L_3 \leftarrow L_3 - 3L_1}} \left( \begin{array}{ccc|c} 1 & 1 & 1 & -1 \\ 0 & 2 & 6 & 2 \\ 0 & 6 & 24 & 4 \end{array} \right) \xrightarrow{\substack{L_1 \leftarrow L_1 - L_2 \\ L_3 \leftarrow L_3 - 6L_2}} \left( \begin{array}{ccc|c} 1 & 0 & -2 & -2 \\ 0 & 1 & 3 & 1 \\ 0 & 0 & 6 & -2 \end{array} \right) \xrightarrow{\substack{L_1 \leftarrow L_1 + 2L_3 \\ L_2 \leftarrow L_2 - 3L_3}} \left( \begin{array}{ccc|c} 1 & 0 & 0 & -8/3 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & -1/3 \end{array} \right)$$

donc  $a_3 = -\frac{1}{3}$ ,  $a_2 = 2$ ,  $a_1 = -\frac{8}{3}$  et  $a_0 = 2$  et on trouve  $P(x) = 2 - \frac{8}{3}x + 2x^2 - \frac{1}{3}x^3$ .

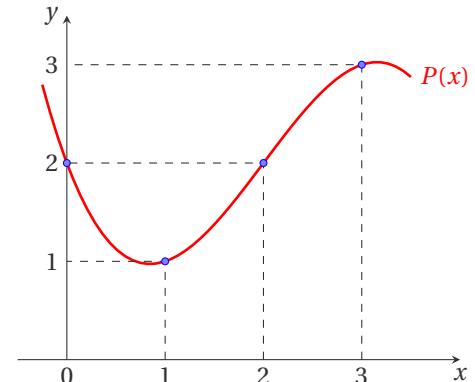
### • Méthode de Lagrange

On considère  $\mathcal{L} = \{L_0, L_1, L_2, L_3\}$  une base de  $\mathbb{R}_3[x]$  telle que  $\text{coord}(P, \mathcal{L}) = (y_0, y_1, y_2, y_3)$ , i.e.  $P(x) = \sum_{i=0}^3 y_i L_i(x)$ . On a

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

donc

$$\begin{aligned} P(x) &= y_0 \frac{(x - x_1)(x - x_2)(x - x_3)}{(x_0 - x_1)(x_0 - x_2)(x_0 - x_3)} + y_1 \frac{(x - x_0)(x - x_2)(x - x_3)}{(x_1 - x_0)(x_1 - x_2)(x_1 - x_3)} \\ &\quad + y_2 \frac{(x - x_0)(x - x_1)(x - x_3)}{(x_2 - x_0)(x_2 - x_1)(x_2 - x_3)} + y_3 \frac{(x - x_0)(x - x_1)(x - x_2)}{(x_3 - x_0)(x_3 - x_1)(x_3 - x_2)} = \\ &= 2 \frac{(x - 1)(x - 2)(x - 3)}{(0 - 1)(0 - 2)(0 - 3)} + \frac{(x - 0)(x - 2)(x - 3)}{(1 - 0)(1 - 2)(1 - 3)} \\ &\quad + 2 \frac{(x - 0)(x - 1)(x - 3)}{(2 - 0)(2 - 1)(2 - 3)} + 3 \frac{(x - 0)(x - 1)(x - 2)}{(3 - 0)(3 - 1)(3 - 2)} = \\ &= \frac{(x - 1)(x - 2)(x - 3)}{-3} + \frac{x(x - 2)(x - 3)}{2} \\ &\quad - x(x - 1)(x - 3) + \frac{x(x - 1)(x - 2)}{2} = -\frac{1}{3}x^3 + 2x^2 - \frac{8}{3}x + 2. \end{aligned}$$



### • Méthode de Newton

On considère  $\mathcal{N} = \{\omega_0, \omega_1, \omega_2, \omega_3\}$  une base de  $\mathbb{R}_3[x]$  telle que  $\text{coord}(P, \mathcal{N}) = (y_0, f[x_0, x_1], f[x_0, x_1, x_2], f[x_0, x_1, x_2, x_3])$ , i.e.  $P(x) = \sum_{i=0}^3 f[x_0, \dots, x_i] \omega_i(x)$ .

La base de Newton est définie récursivement comme suit :

$$\omega_0(x) = 1; \quad \text{pour } k = 1, \dots, n \quad \omega_k(x) = \omega_{k-1}(x)(x - x_{k-1}).$$

Les coordonnées sont les valeurs encadrées dans le tableau des différences divisées ci-dessous :

$i$	$x_i$	$y_i$	$f[x_{i-1}, x_i]$	$f[x_{i-2}, x_{i-1}, x_i]$	$f[x_{i-3}, x_{i-2}, x_{i-1}, x_i]$
0	0	2			
1	1	1	-1		
2	2	2	1	1	
3	3	3	1	0	- $\frac{1}{3}$

On a alors

$$\begin{aligned}
 P_3(x) &= \sum_{i=1}^3 f[x_0, \dots, x_i] \omega_i(x) \\
 &= y_0 \omega_0(x) + f[x_0, x_1] \omega_1(x) + f[x_0, x_1, x_2] \omega_2(x) + f[x_0, x_1, x_2, x_3] \omega_3(x) \\
 &= 2\omega_0(x) - \omega_1(x) + \omega_2(x) - \frac{1}{3}\omega_3(x) \\
 &= 2 - x + x(x-1) - \frac{1}{3}x(x-1)(x-2) \\
 &= -\frac{1}{3}x^3 + 2x^2 - \frac{8}{3}x + 2.
 \end{aligned}$$

Remarque : on réordonne les points comme suit : (1, 1), (2, 2), (3, 3) et (0, 2).

$i$	$x_i$	$y_i$	$f[x_{i-1}, x_i]$	$f[x_{i-2}, x_{i-1}, x_i]$	$f[x_{i-3}, x_{i-2}, x_{i-1}, x_i]$
0	1	1			
1	2	2	1		
2	3	3	1	0	
3	0	2	$-\frac{1}{3}$	$-\frac{1}{3}$	$-\frac{1}{3}$

On a alors

$$\begin{aligned}
 P_3(x) &= \sum_{i=1}^3 f[x_0, \dots, x_i] \omega_i(x) \\
 &= y_0 \omega_0(x) + f[x_0, x_1] \omega_1(x) + f[x_0, x_1, x_2] \omega_2(x) + f[x_0, x_1, x_2, x_3] \omega_3(x) \\
 &= \omega_0(x) + \omega_1(x) - \frac{1}{3}\omega_3(x) \\
 &= 1 + (x-1) - \frac{1}{3}(x-1)(x-2)(x-3) \\
 &= x - \frac{1}{3}(x-1)(x-2)(x-3).
 \end{aligned}$$

On remarque que les points (1, 1), (2, 2) et (3, 3) sont alignés, ainsi le polynôme  $Q(x) = x$  de  $\mathbb{R}_2[x]$  interpole ces points.

### Exercice 2.3

1. Calculer le polynôme d'interpolation de la fonction  $f(x) = \cos(x)$  en les 3 points  $x_i = \frac{\pi}{2}i$  avec  $i = 0, \dots, 2$ .
2. Calculer ensuite le polynôme d'interpolation de la même fonction en les 4 points  $x_i = \frac{\pi}{2}i$  avec  $i = 0, \dots, 3$ , i.e. en ajoutant le point  $x_3 = 3\pi/2$ .

### Correction

1. On cherche  $p_2 \in \mathbb{R}_2[x]$  tel que  $p_2(x_i) = \cos(x_i)$  pour  $i = 0, \dots, 2$ . On peut choisir l'une des quatre méthodes ci-dessous (on préférera la méthode de NEWTON car elle permet de réutiliser les calculs de cette question pour répondre à la question suivante).

**Méthode directe (naïve).** Si on écrit  $p_2(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2$ , on cherche  $\alpha_0, \alpha_1, \alpha_2$  tels que

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & \frac{\pi}{2} & \frac{\pi^2}{4} \\ 1 & \pi & \pi^2 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}$$

En résolvant ce système linéaire on trouve  $\alpha_0 = 1$ ,  $\alpha_1 = -\frac{2}{\pi}$  et  $\alpha_2 = 0$  :

```
Px=[0:1:2]*pi/2
Py=cos(Px);
P=[Px' Py']
alpha=naivePoly(P)
```

**Méthode astucieuse.** Le polynôme  $p_2$  s'annule en  $\frac{\pi}{2}$ , ceci signifie qu'il existe un polynôme  $R(x)$  tel que

$$p_2(x) = R(x) \left( x - \frac{\pi}{2} \right).$$

Puisque  $p_2(x)$  a degré 2, le polynôme  $R(x)$  qu'on a mis en facteur a degré 1, autrement dit  $R$  est de la forme  $ax + b$ . On cherche alors  $a$  et  $b$  tels que

$$\begin{cases} R(0) = \frac{p_2(0)}{(0 - \frac{\pi}{2})}, \\ R(\pi) = \frac{p_2(\pi)}{(\pi - \frac{\pi}{2})}. \end{cases} \iff \begin{cases} b = \frac{1}{(0 - \frac{\pi}{2})}, \\ a\pi + b = \frac{-1}{(\pi - \frac{\pi}{2})}. \end{cases} \iff \begin{cases} b = -\frac{2}{\pi}, \\ a = 0. \end{cases}$$

Ainsi

$$p_2(x) = R(x) \left( x - \frac{\pi}{2} \right) = -\frac{2}{\pi} \left( x - \frac{\pi}{2} \right) = -\frac{2}{\pi} x + 1.$$

**Méthode de Lagrange.** On a

$$p_2(x) = y_0 L_0(x) + y_1 L_1(x) + y_2 L_2(x) = 1 \frac{(x - \frac{\pi}{2})(x - \pi)}{(0 - \frac{\pi}{2})(0 - \pi)} - 1 \frac{(x - 0)(x - \frac{\pi}{2})}{(\pi - 0)(\pi - \frac{\pi}{2})} = 1 - \frac{2}{\pi} x.$$

**Méthode de Newton.** On commence par construire le tableau des différences divisées :

$i$	$x_i$	$y_i$	$f[x_{i-1}, x_i]$	$f[x_{i-2}, x_{i-1}, x_i]$
0	0	1		
1	$\frac{\pi}{2}$	0	$-\frac{2}{\pi}$	
2	$\pi$	-1	$-\frac{2}{\pi}$	0

On a alors

$$\begin{aligned} p_2(x) &= \sum_{i=1}^2 \omega_i(x) f[x_0, \dots, x_i] \\ &= \omega_0(x) f[x_0] + \omega_1(x) f[x_0, x_1] + \omega_2(x) f[x_0, x_1, x_2] \\ &= \omega_0(x) - \frac{2}{\pi} \omega_1(x) \\ &= 1 - \frac{2}{\pi} x. \end{aligned}$$

2. On cherche donc  $p_3 \in \mathbb{R}_3[x]$  tel que  $p_3(x_i) = \sin(x_i)$  pour  $i = 0, \dots, 3$ . On peut choisir l'une des quatre méthodes ci-dessous (on préférera la méthode de NEWTON car elle permet d'utiliser les calculs précédents).

**Méthode directe.** Si on écrit  $p_3(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3$ , on cherche  $\alpha_0, \alpha_1, \alpha_2, \alpha_3$  tels que

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & \frac{\pi}{2} & \frac{\pi^2}{4} & \frac{\pi^3}{8} \\ 1 & \pi & \pi^2 & \pi^3 \\ 1 & \frac{3\pi}{2} & \frac{9\pi^2}{4} & \frac{27\pi^3}{8} \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ -1 \\ 0 \end{pmatrix}$$

En résolvant ce système linéaire on trouve  $\alpha_0 = 1$ ,  $\alpha_1 = -\frac{2}{3\pi}$ ,  $\alpha_2 = -\frac{4}{\pi^2}$  et  $\alpha_3 = \frac{8}{3\pi^3}$ :

```
Px=[0:1:3]*pi/2
Py=cos(Px);
P=[Px' Py']
alpha=naivePoly(P)
```

**Méthode astucieuse.** Le polynôme  $p_3$  s'annule en  $\frac{\pi}{2}$  et en  $\frac{3\pi}{2}$ , ceci signifie qu'il existe un polynôme  $R(x)$  tel que

$$p_3(x) = R(x) \left( x - \frac{\pi}{2} \right) \left( x - \frac{3\pi}{2} \right).$$

Puisque  $p_3(x)$  a degré 3, le polynôme  $R(x)$  qu'on a mis en facteur a degré 1, autrement dit  $R$  est de la forme  $ax + b$ . On cherche alors  $a$  et  $b$  tels que

$$\begin{cases} R(0) = \frac{p_3(0)}{(0 - \frac{\pi}{2})(0 - \frac{3\pi}{2})}, \\ R(\pi) = \frac{p_3(\pi)}{(\pi - \frac{\pi}{2})(\pi - \frac{3\pi}{2})}. \end{cases} \iff \begin{cases} b = \frac{1}{(0 - \frac{\pi}{2})(0 - \frac{3\pi}{2})}, \\ a\pi + b = \frac{-1}{(\pi - \frac{\pi}{2})(\pi - \frac{3\pi}{2})}. \end{cases} \iff \begin{cases} b = \frac{4}{3\pi^2}, \\ a = \frac{8}{3\pi^3}. \end{cases}$$

Ainsi

$$p_3(x) = R(x) \left( x - \frac{\pi}{2} \right) \left( x - \frac{3\pi}{2} \right) = \left( \frac{8}{3\pi^3} x + \frac{4}{3\pi^2} \right) \left( x - \frac{\pi}{2} \right) \left( x - \frac{3\pi}{2} \right) = 1 - \frac{2}{3\pi} x - \frac{4}{\pi^2} x^2 + \frac{8}{3\pi^3} x^3.$$

**Méthode de Lagrange.** On a

$$\begin{aligned} p_3(x) &= y_0 L_0(x) + y_1 L_1(x) + y_2 L_2(x) + y_3 L_3(x) = 1 \frac{(x - \frac{\pi}{2})(x - \pi)(x - \frac{3\pi}{2})}{(0 - \frac{\pi}{2})(0 - \pi)(0 - \frac{3\pi}{2})} - 1 \frac{(x - 0)(x - \frac{\pi}{2})(x - \frac{3\pi}{2})}{(\pi - 0)(\pi - \frac{\pi}{2})(\pi - \frac{3\pi}{2})} \\ &= \frac{4}{3\pi^3} \left( x - \frac{\pi}{2} \right) \left( x - \frac{3\pi}{2} \right) (-x + \pi + 3x) = 1 - \frac{2}{3\pi} x - \frac{4}{\pi^2} x^2 + \frac{8}{3\pi^3} x^3. \end{aligned}$$

**Méthode de Newton.** Il suffit de calculer une différence divisée en plus, i.e. ajouter une ligne au tableau précédent :

$i$	$x_i$	$y_i$	$f[x_{i-1}, x_i]$	$f[x_{i-2}, x_{i-1}, x_i]$	$f[x_{i-3}, x_{i-2}, x_{i-1}, x_i]$
0	0	1			
1	$\frac{\pi}{2}$	0	$-\frac{2}{\pi}$		
2	$\pi$	-1	$-\frac{2}{\pi}$	0	
3	$\frac{3\pi}{2}$	0	$\frac{2}{\pi}$	$\frac{4}{\pi^2}$	$\frac{8}{3\pi^3}$

On a alors

$$\begin{aligned} p_3(x) &= \sum_{i=1}^3 \omega_i(x) f[x_0, \dots, x_i] \\ &= p_2(x) + \omega_3(x) f[x_0, x_1, x_2, x_3] \\ &= 1 - \frac{2}{\pi} x + \frac{8}{3\pi^3} \omega_3(x) \\ &= 1 - \frac{2}{\pi} x + \frac{8}{3\pi^3} x \left( x - \frac{\pi}{2} \right) (x - \pi) \\ &= 1 - \frac{2}{3\pi} x - \frac{4}{\pi^2} x^2 + \frac{8}{3\pi^3} x^3. \end{aligned}$$

### Exercice 2.4

- Construire le polynôme les points  $(-1, 2), (0, 1), (1, 2)$  et  $(2, 3)$ .
- Soit  $Q$  le polynôme qui interpole les points  $(-1, 2), (0, 1), (1, 2)$ . Montrer qu'il existe un réel  $\lambda$  tel que :

$$Q(x) - P(x) = \lambda(x+1)x(x-1).$$

### Correction

- Dans la base de LAGRANGE le polynôme d'interpolation de degré  $n = 3$  s'écrit

$$\begin{aligned} P(x) &= y_0 \frac{(x - x_1)(x - x_2)(x - x_3)}{(x_0 - x_1)(x_0 - x_2)(x_0 - x_3)} + y_1 \frac{(x - x_0)(x - x_2)(x - x_3)}{(x_1 - x_0)(x_1 - x_2)(x_1 - x_3)} \\ &\quad + y_2 \frac{(x - x_0)(x - x_1)(x - x_3)}{(x_2 - x_0)(x_2 - x_1)(x_2 - x_3)} + y_3 \frac{(x - x_0)(x - x_1)(x - x_2)}{(x_3 - x_0)(x_3 - x_1)(x_3 - x_2)} \\ &= \frac{x(x-1)(x-2)}{-3} + \frac{(x+1)(x-1)(x-2)}{2} - (x+1)x(x-2) + \frac{(x+1)x(x-1)}{2} = \\ &= -\frac{1}{3}x^3 + x^2 + \frac{1}{3}x + 1. \end{aligned}$$

```
P=[-1 2; 0 1; 1 2; 2 3];
alpha=naivePoly(P)
```

## 2. Par construction

$$\begin{aligned} Q(-1) &= P(-1), \\ Q(0) &= P(0), \\ Q(1) &= P(1), \end{aligned}$$

donc le polynôme  $Q(x) - P(x)$  s'annule en  $-1$ , en  $0$  et en  $1$ , ceci signifie qu'il existe un polynôme  $R(x)$  tel que

$$Q(x) - P(x) = R(x)(x+1)x(x-1).$$

Puisque  $P(x)$  a degré 3 et  $Q(x)$  a degré 2, le polynôme  $Q(x) - P(x)$  a degré 3, donc le polynôme  $R(x)$  qu'on a mis en facteur a degré 0 (*i.e.*  $R(x)$  est une constante).

Si on n'a pas remarqué ça, on peut tout de même faire tous les calculs : dans ce cas  $n = 2$  donc on a

$$\begin{aligned} Q(x) &= y_0 \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} + y_1 \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} + y_2 \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} \\ &= x(x-1) - (x+1)(x-1) + (x+1)x \\ &= x^2 + 1. \end{aligned}$$

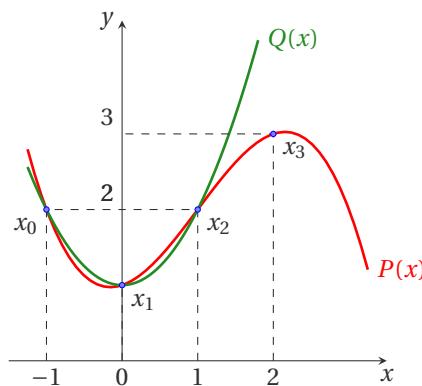
Ainsi

$$\begin{aligned} Q(x) - P(x) &= y_0 \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} \left[ 1 - \frac{x-x_3}{x_0-x_3} \right] + y_1 \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} \left[ 1 - \frac{x-x_3}{x_1-x_3} \right] \\ &\quad + y_2 \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} \left[ 1 - \frac{x-x_3}{x_2-x_3} \right] - y_3 \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} \\ &= -y_0 \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} - y_1 \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} \\ &\quad - y_2 \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} - y_3 \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} \\ &= - \left[ \frac{y_0}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} + \frac{y_1}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} \right. \\ &\quad \left. + \frac{y_2}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} + \frac{y_3}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} \right] (x-x_0)(x-x_1)(x-x_2) \\ &= \frac{(x+1)x(x-1)}{3} \end{aligned}$$

et  $\lambda = \frac{1}{3}$ . Sinon directement

$$Q(x) - P(x) = x^2 + 1 + \frac{1}{3}x^3 - x^2 + \frac{1}{3}x - 1 = \frac{1}{3}x^3 + \frac{1}{3}x = \frac{(x+1)x(x-1)}{3} = \lambda x(x+1)(x-1)$$

avec  $\lambda = \frac{1}{3}$ .



### Exercice 2.5

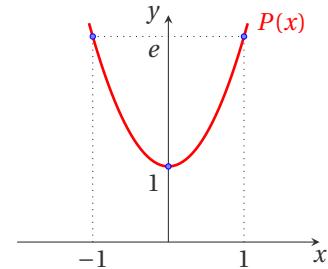
- Construire le polynôme  $P$  qui interpolate les trois points  $(-1, e)$ ,  $(0, 1)$  et  $(1, e)$ .

2. Sans faire de calculs, donner l'expression du polynôme  $Q$  qui interpole les trois points  $(-1, -1)$ ,  $(0, 0)$  et  $(1, -1)$ .
3. Trouver le polynôme de l'espace vectoriel  $\text{Vec}\{1, x, x^2\}$  qui interpole les trois points  $(-1, -1)$ ,  $(0, 0)$  et  $(1, -1)$ .

**Correction**

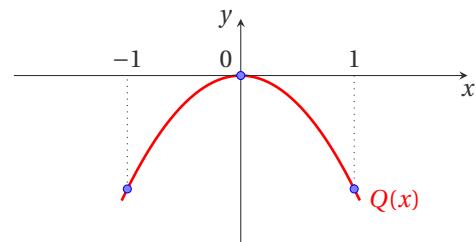
1. Dans la base de LAGRANGE le polynôme d'interpolation de degré  $n = 2$  s'écrit

$$\begin{aligned} P(x) &= y_0 \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} + y_1 \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} + y_2 \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} = \\ &= e \frac{x(x-1)}{2} - (x+1)(x-1) + e \frac{(x+1)x}{2} = \\ &= (e-1)x^2 + 1. \end{aligned}$$



2. Il suffit de changer les coefficients  $y_i$  dans l'expression précédente :

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



3. Il s'agit de trouver un polynôme  $p(x)$  qui soit combinaison linéaire des deux polynômes assignés (i.e.  $p(x) = \alpha + \beta x + \gamma x^2$ ) et qui interpole les trois points  $(-1, -1)$ ,  $(0, 0)$  et  $(1, -1)$  :

$$\begin{cases} p(-1) = 1, \\ p(0) = 0, \\ p(1) = -1, \end{cases} \Leftrightarrow \begin{cases} \alpha - \beta + \gamma = -1, \\ \alpha = 0, \\ \alpha + \beta + \gamma = -1, \end{cases}$$

d'où  $\alpha = 0$ ,  $\beta = 0$  et  $\gamma = -1$ . Le polynôme cherché est donc le polynôme  $p(x) = -x^2$ . En fait, il suffisait de remarquer que le polynôme  $Q \in \text{Vec}\{1, x, x^2\}$  pour conclure que le polynôme  $p$  cherché est  $Q$  lui-même.

**Exercice 2.6**

1. Construire le polynôme  $P$  qui interpole les points  $(-1, 1)$ ,  $(0, 1)$ ,  $(1, 2)$  et  $(2, 3)$ .
2. Soit  $Q$  le polynôme qui interpole les points  $(-1, 1)$ ,  $(0, 1)$ ,  $(1, 2)$ . Montrer qu'il existe un réel  $\lambda$  tel que :

$$Q(x) - P(x) = \lambda(x+1)x(x-1).$$

**Correction**

1. Dans la base de LAGRANGE le polynôme d'interpolation de degré  $n = 3$  s'écrit

$$\begin{aligned} P(x) &= y_0 \frac{(x - x_1)(x - x_2)(x - x_3)}{(x_0 - x_1)(x_0 - x_2)(x_0 - x_3)} + y_1 \frac{(x - x_0)(x - x_2)(x - x_3)}{(x_1 - x_0)(x_1 - x_2)(x_1 - x_3)} \\ &\quad + y_2 \frac{(x - x_0)(x - x_1)(x - x_3)}{(x_2 - x_0)(x_2 - x_1)(x_2 - x_3)} + y_3 \frac{(x - x_0)(x - x_1)(x - x_2)}{(x_3 - x_0)(x_3 - x_1)(x_3 - x_2)} \\ &= \frac{x(x-1)(x-2)}{-6} + \frac{(x+1)(x-1)(x-2)}{2} - (x+1)x(x-2) + \frac{(x+1)x(x-1)}{2} = \\ &= -\frac{1}{6}x^3 + \frac{1}{2}x^2 + \frac{2}{3}x + 1. \end{aligned}$$

2. Par construction

$$\begin{aligned} Q(-1) &= P(-1), \\ Q(0) &= P(0), \\ Q(1) &= P(1), \end{aligned}$$

donc le polynôme  $Q(x) - P(x)$  s'annule en  $-1$ , en  $0$  et en  $1$ , ceci signifie qu'il existe un polynôme  $R(x)$  tel que

$$Q(x) - P(x) = R(x)(x+1)x(x-1).$$

Puisque  $P(x)$  a degré 3 et  $Q(x)$  a degré 2, le polynôme  $Q(x) - P(x)$  a degré 3, donc le polynôme  $R(x)$  qu'on a mis en facteur a degré 0 (*i.e.*  $R(x)$  est une constante).

Si on n'a pas remarqué ça, on peut tout de même faire tous les calculs : dans ce cas  $n = 2$  donc on a

$$\begin{aligned} Q(x) &= y_0 \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} + y_1 \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} + y_2 \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} \\ &= \frac{x(x-1)}{2} - (x+1)(x-1) + (x+1)x \\ &= \frac{1}{2}x^2 + \frac{1}{2}x + 1. \end{aligned}$$

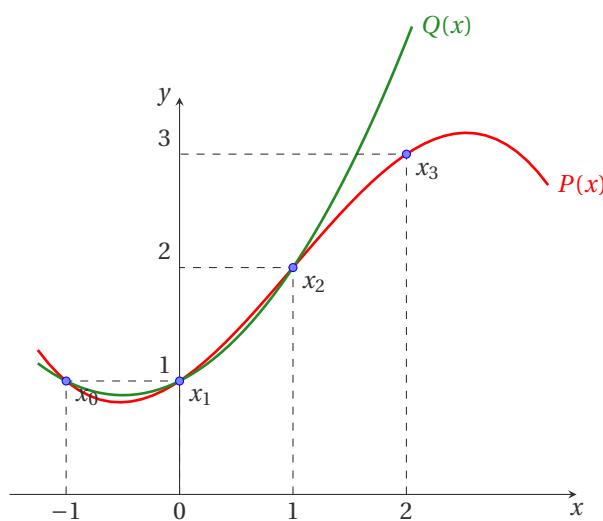
Ainsi

$$\begin{aligned} Q(x) - P(x) &= y_0 \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} \left[ 1 - \frac{x-x_3}{x_0-x_3} \right] + y_1 \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} \left[ 1 - \frac{x-x_3}{x_1-x_3} \right] \\ &\quad + y_2 \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} \left[ 1 - \frac{x-x_3}{x_2-x_3} \right] - y_3 \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} \\ &= -y_0 \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} - y_1 \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} \\ &\quad - y_2 \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} - y_3 \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} \\ &= - \left[ \frac{y_0}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} + \frac{y_1}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} \right. \\ &\quad \left. + \frac{y_2}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} + \frac{y_3}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} \right] (x-x_0)(x-x_1)(x-x_2) = \frac{(x+1)x(x-1)}{6} \end{aligned}$$

et  $\lambda = \frac{1}{6}$ . Sinon directement

$$Q(x) - P(x) = \frac{1}{2}x^2 + \frac{1}{2}x + 1 + \frac{1}{6}x^3 - \frac{1}{2}x^2 - \frac{2}{3}x - 1 = \frac{1}{6}x^3 - \frac{1}{6}x = \frac{1}{6}x(x^2 - 1) = \lambda x(x+1)(x-1)$$

avec  $\lambda = \frac{1}{6}$ .



### Exercice 2.7

1. Construire le polynôme  $P$  qui interpolate les trois points  $(-1, \alpha)$ ,  $(0, \beta)$  et  $(1, \alpha)$  où  $\alpha$  et  $\beta$  sont des réels.
2. Si  $\alpha = \beta$ , donner le degré de  $P$ .
3. Montrer que  $P$  est pair. Peut-on avoir  $P$  de degré 1 ?

**Correction**

1. Dans la base de LAGRANGE le polynôme d'interpolation de degré  $n = 2$  s'écrit

$$\begin{aligned} P(x) &= y_0 \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} + y_1 \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} + y_2 \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} + \\ &= \alpha \frac{x(x-1)}{2} + \beta \frac{(x+1)(x-1)}{-1} + \gamma \frac{(x+1)x}{2} = \\ &= \frac{\alpha}{2} x(x-1) - \beta(x+1)(x-1) + \frac{\alpha}{2} x(x+1) \\ &= (\alpha - \beta)x^2 + \beta. \end{aligned}$$

Sinon, dans la base de NEWTON, on commence par construire le tableau des différences divisées :

$i$	$x_i$	$y_i$	$f[x_{i-1}, x_i]$	$f[x_{i-2}, x_{i-1}, x_i]$
0	-1	$\alpha$		
1	0	$\beta$	$(\beta - \alpha)$	
2	1	$\alpha$	$(\alpha - \beta)$	$(\alpha - \beta)$

On a alors

$$\begin{aligned} p_2(x) &= \sum_{i=0}^2 \omega_i(x) f[x_0, \dots, x_i] \\ &= \omega_0(x) f[x_0] + \omega_1(x) f[x_0, x_1] + \omega_2(x) f[x_0, x_1, x_2] \\ &= \alpha \omega_0(x) + (\beta - \alpha) \omega_1(x) + (\alpha - \beta) \omega_2(x) \\ &= \alpha + (\beta - \alpha)(x+1) + (\alpha - \beta)x(x+1) = (\alpha - \beta)x^2 + \beta. \end{aligned}$$

2. Si  $\alpha = \beta$ ,  $P(x) = \alpha$  qui est un polynôme de degré 0.

3.  $P(-x) = P(x)$  donc  $P$  est pair. Donc  $P$  ne peut pas être de degré 1 car un polynôme de degré 1 est de la forme  $a_0 + a_1 x$  qui ne peut pas être pair.

**Exercice 2.8**

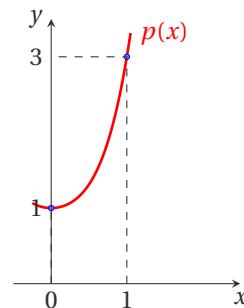
Trouver le polynôme de l'espace vectoriel  $\text{Vec}\{1 + x^2, x^4\}$  qui interpole les points  $(0, 1)$  et  $(1, 3)$ .

**Correction**

Il s'agit de trouver un polynôme  $p(x)$  qui soit combinaison linéaire des deux polynômes assignés (i.e.  $p(x) = \alpha(1 + x^2) + \beta(x^4)$ ) et qui interpole les deux points  $(0, 1)$  et  $(1, 3)$  :

$$\begin{cases} p(0) = 1, \\ p(1) = 3, \end{cases} \Leftrightarrow \begin{cases} \alpha(1 + 0^2) + \beta(0^4) = 1, \\ \alpha(1 + 1^2) + \beta(1^4) = 3, \end{cases}$$

d'où  $\alpha = 1$  et  $\beta = 1$ . Le polynôme cherché est donc le polynôme  $p(x) = 1 + x^2 + x^4$ .

**Exercice 2.9**

Soit  $f: \mathbb{R} \rightarrow \mathbb{R}$  la fonction définie par  $f(x) = 1 + x^3$ .

1. Calculer le polynôme  $p_0 \in \mathbb{R}_0[x]$  qui interpole  $f$  au point d'abscisse  $x_0 = 0$ .
2. Calculer le polynôme  $p_1 \in \mathbb{R}_1[x]$  qui interpole  $f$  aux points d'abscisse  $\{x_0 = 0, x_1 = 1\}$ .
3. Calculer le polynôme  $p_2 \in \mathbb{R}_2[x]$  qui interpole  $f$  aux points d'abscisse  $\{x_0 = 0, x_1 = 1, x_2 = 2\}$ .
4. Calculer le polynôme  $p_3 \in \mathbb{R}_3[x]$  qui interpole  $f$  aux points d'abscisse  $\{x_0 = 0, x_1 = 1, x_2 = 2, x_3 = 3\}$ .
5. Pour  $n > 3$ , calculer les polynômes  $p_n \in \mathbb{R}_n[x]$  qui interpolent  $f$  aux points d'abscisse  $\{x_0 = 0, x_1 = 1, \dots, x_n = n\}$ .

**Correction**

1. On interpole l'ensemble  $\{(0, 1)\}$  donc  $p_0(x) = 1$ .
2. On interpole l'ensemble  $\{(0, 1), (1, 2)\}$  donc  $p_1(x) = 1 + x$ .
3. On interpole l'ensemble  $\{(0, 1), (1, 2), (2, 9)\}$  donc  $p_2(x) = 1 - 2x + 3x^2$ .
4.  $f \in \mathbb{R}_3[x]$  et comme il existe un seul polynôme de degré au plus 3 qui interpole quatre points ce polynôme coïncide forcément avec  $f$  donc  $p_3 \equiv f$ .
5.  $f \in \mathbb{R}_n[x]$  pour tout  $n \geq 3$  et comme il existe un seul polynôme de degré au plus 3 qui interpole quatre points ce polynôme coïncide forcément avec  $f$  donc  $p_n \equiv f$  pour  $n \geq 3$ .

**Exercice 2.10**

Soit  $f : \mathbb{R} \rightarrow \mathbb{R}$  la fonction définie par  $f(x) = 1 + x^2$ .

1. Calculer le polynôme de  $\mathbb{R}_0[x]$  qui interpole  $f$  au point 0.
2. Calculer le polynôme de  $\mathbb{R}_1[x]$  qui interpole  $f$  aux points  $\{0, 2\}$ .
3. Calculer le polynôme de  $\mathbb{R}_9[x]$  qui interpole  $f$  aux points  $\{0, 2, \dots, 2i, \dots, 18\}_{0 \leq i \leq 9}$ .

**Correction**

1. On interpole l'ensemble de points  $\{(0, 1)\}$  donc  $p_0(x) = 1$ .
2. On interpole l'ensemble de points  $\{(0, 1), (2, 5)\}$  donc  $p_1(x) = 1 + 2x$ .
3.  $f \in \mathbb{R}_n[x]$  pour tout  $n \geq 2$  et comme il existe un seul polynôme de degré au plus 2 qui interpole trois points ce polynôme coïncide forcément avec  $f$  donc  $p_n \equiv f$  pour  $n \geq 2$ .

**Exercice 2.11**

1. Calculer le polynôme qui interpole les points  $(0, 3), (1, 2), (2, 4), (3, -2)$ .
2. Calculer le polynôme qui interpole les points  $(0, 2), (1, 3), (2, 4), (3, 5), (4, 6), (5, 7), (6, 8), (7, 9)$  (pas de calculs inutiles!).
3. Calculer le polynôme qui interpole les points  $(0, 2), (1, 1), (2, 2), (3, 3), (4, 4)$  en le cherchant sous la forme  $p(x) = x + q(x)$  (pas de calculs inutiles!).
4. Donner l'expression du polynôme  $p \in \mathbb{R}_3[x]$  dont la dérivée  $k$ -ème vérifie  $p^{(k)}(1) = 3$  pour  $k = 0, 1, 2, 3$ . Est-il unique dans  $p \in \mathbb{R}_3[x]$ ? Soit  $f$  une fonction de classe  $\mathcal{C}^\infty$  telle que  $f^{(k)}(1) = 3$ . Quelle estimation de  $f(x) - p(x)$  a-t-on?

**Correction**

1. Dans la base de LAGRANGE le polynôme d'interpolation de degré  $n = 3$  s'écrit

$$\begin{aligned} P(x) &= y_0 \frac{(x - x_1)(x - x_2)(x - x_3)}{(x_0 - x_1)(x_0 - x_2)(x_0 - x_3)} + y_1 \frac{(x - x_0)(x - x_2)(x - x_3)}{(x_1 - x_0)(x_1 - x_2)(x_1 - x_3)} \\ &\quad + y_2 \frac{(x - x_0)(x - x_1)(x - x_3)}{(x_2 - x_0)(x_2 - x_1)(x_2 - x_3)} + y_3 \frac{(x - x_0)(x - x_1)(x - x_2)}{(x_3 - x_0)(x_3 - x_1)(x_3 - x_2)} \\ &= 3 \frac{(x - 1)(x - 2)(x - 3)}{(0 - 1)(0 - 2)(0 - 3)} + 2 \frac{x(x - 2)(x - 3)}{(1 - 0)(1 - 2)(1 - 3)} \\ &\quad + 4 \frac{x(x - 1)(x - 3)}{(2 - 0)(2 - 1)(2 - 3)} - 2 \frac{x(x - 1)(x - 2)}{(3 - 0)(3 - 1)(3 - 2)} \\ &= 3 - \frac{37}{6}x + 7x^2 - \frac{11}{6}x^3. \end{aligned}$$

2.  $p(x) = x + 2$ : en effet, on voit que les points sont alignés le long de la droite d'équation  $y = x + 2$ .
3.  $p \in \mathbb{R}_4[x]$  et interpole les points  $(0, 2), (1, 1), (2, 2), (3, 3), (4, 4)$  donc  $p(0) = 2, p(1) = 1, p(2) = 2, p(3) = 3$  et  $p(4) = 4$ .
  - \* Première méthode. On cherche le polynôme  $q \in \mathbb{R}_4[x]$  tel que  $q(x) = p(x) - x$ , autrement dit le polynôme  $q \in \mathbb{R}_4[x]$  qui interpole les points  $(0, 2 - 0), (1, 1 - 1), (2, 2 - 2), (3, 3 - 3), (4, 4 - 4)$ . Donc le polynôme  $q$  s'annule en  $x = 1, x = 2, x = 3$  et  $x = 4$ , ceci signifie qu'il existe un polynôme  $R$  tel que

$$q(x) = (x - 1)(x - 2)(x - 3)(x - 4)R(x).$$

Comme  $q \in \mathbb{R}_4[x]$  alors  $R$  est une constante qu'on peut calculer en imposant  $q(0) = 2$  et l'on obtient

$$q(x) = \frac{1}{12}(x - 1)(x - 2)(x - 3)(x - 4).$$

- \* **Deuxième méthode.** Notons  $x_0 = 1, x_1 = 2, x_2 = 3, x_3 = 4, x_4 = 0$ . On considère  $\mathcal{N} = \{\omega_0, \omega_1, \omega_2, \dots, \omega_{n-1}\}$  une base de  $\mathbb{R}_{n-1}[x]$  telle que  $\text{coord}(p, \mathcal{N}) = (y_0, f[x_0, x_1], f[x_0, x_1, x_2], \dots, f[x_0, x_1, x_2, \dots, x_{n-1}])$ , i.e.  $p(x) = \sum_{i=0}^{n-1} f[x_0, \dots, x_i] \omega_i(x)$ . La base de NEWTON est définie récursivement comme suit :

$$\omega_0(x) = 0; \quad \omega_1(x) = x - x_0; \quad \text{pour } k = 2, \dots, n \quad \omega_k(x) = \omega_{k-1}(x)(x - x_{k-1}).$$

Le polynôme d'interpolation de degré  $n$  sur l'ensemble des  $n+1$  points  $\{(x_i, y_i)\}_{i=1}^n$  dans la base de NEWTON s'écrit

$$\begin{aligned} p_n(x) &= \sum_{i=0}^n \omega_i(x) f[x_0, \dots, x_i] \\ &= \sum_{i=1}^{n-1} \omega_i(x) f[x_0, \dots, x_i] + \omega_4(x) f[x_0, x_1, x_2, x_3, x_4] \\ &= p_{n-1}(x) + \omega_4(x) f[x_0, x_1, x_2, x_3, x_4] \end{aligned}$$

où  $p_{n-1}$  est le polynôme d'interpolation de degré  $n-1$  sur l'ensemble des  $n$  points  $\{(x_i, y_i)\}_{i=0}^{n-1}$ .

Dans notre cas, on voit que les points  $\{x_0, x_1, x_2, x_3\}$  sont alignés le long de la droite d'équation  $y = x$  donc  $p_{n-1}(x) = x$  et  $q(x) = \omega_4(x) f[x_0, x_1, x_2, x_3, x_4]$  avec  $\omega_4(x) = (x-1)(x-2)(x-3)(x-4)$ . On doit donc calculer le coefficient  $f[x_0, x_1, x_2, x_3, x_4]$  sachant que  $q(0) = 2$ , ce qui donne  $f[x_0, x_1, x_2, x_3, x_4] = 1/12$ .

On conclut que

$$p(x) = x + \frac{1}{12}(x-1)(x-2)(x-3)(x-4) = \frac{1}{12}x^4 - \frac{5}{6}x^3 + \frac{35}{12}x^2 - \frac{19}{6}x + 2.$$

4. Soit  $p(x) = a_0 + a_1x + a_2x^2 + a_3x^3$  un polynôme de  $\mathbb{R}_3[x]$ . On cherche les quatre coefficients  $a_i$  tels que  $p^{(k)}(1) = 3$  pour  $k = 0, 1, 2, 3$  :

$$\begin{cases} p(x) = a_0 + a_1x + a_2x^2 + a_3x^3, \\ p'(x) = a_1 + 2a_2x + 3a_3x^2, \\ p''(x) = 2a_2 + 6a_3x, \\ p'''(x) = 6a_3, \end{cases} \implies \begin{cases} 3 = p(1) = a_0 + a_1 + a_2 + a_3, \\ 3 = p'(1) = a_1 + 2a_2 + 3a_3, \\ 3 = p''(1) = 2a_2 + 6a_3, \\ 3 = p'''(1) = 6a_3, \end{cases} \implies \begin{cases} a_3 = 1/2, \\ a_2 = 0, \\ a_1 = 3/2, \\ a_0 = 1. \end{cases}$$

et ce polynôme est unique.

Soit  $f$  une fonction de classe  $\mathcal{C}^\infty$  telle que  $f^{(k)}(1) = 3$ . Alors la fonction  $g(x) \equiv f(x) - p(x)$  est de classe  $\mathcal{C}^\infty$  et  $g^{(k)}(1) = 0$  pour  $k = 0, 1, 2, 3$  (i.e.  $x = 1$  est un zéro de multiplicité 4 pour  $g$ ). Écrivons le développement de TAYLOR avec le reste de LAGRANGE de  $g$  en  $x = 1$  à l'ordre 3 :

$$g(x) = \sum_{k=0}^3 \frac{g^{(k)}(1)}{k!} (x-1)^k + \frac{g^{(4)}(\xi)}{4!} (x-1)^4 = \frac{g^{(4)}(\xi)}{4!} (x-1)^4$$

où  $\xi$  est entre  $x$  et 1. Le polynôme  $p$  étant de degré 3, on obtient

$$f(x) - p(x) = \frac{f^{(4)}(\xi)}{4!} (x-1)^4.$$

## ★ Exercice 2.12 (Les défauts de l'interpolation polynomiale)

L'objet de cet exercice est de donner un exemple concret de fonction analytique  $f$  pour laquelle les polynômes d'interpolation ne forment pas une suite convergente.

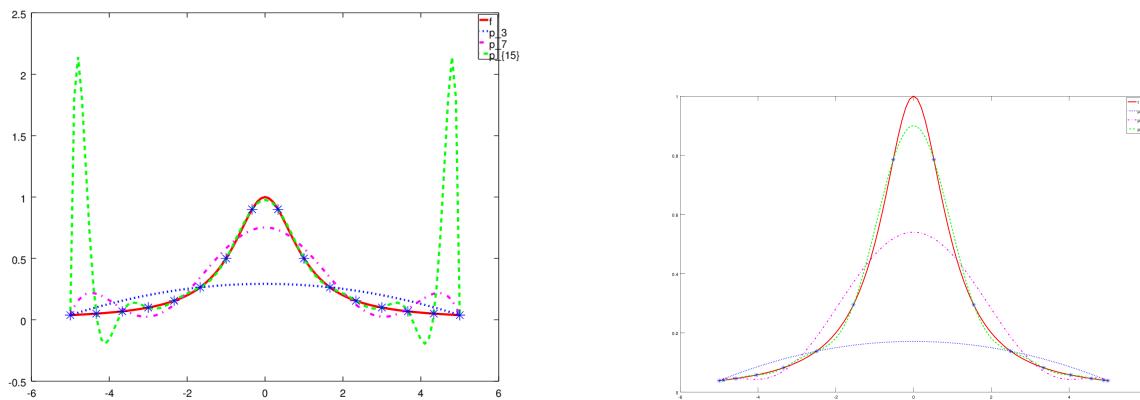
Si  $y_i = f(x_i)$  pour  $i = 1, 2, \dots, n$ ,  $f: I \rightarrow \mathbb{R}$  étant une fonction donnée de classe  $\mathcal{C}^n(I)$  où  $I$  est le plus petit intervalle contenant les noeuds distincts  $\{x_i\}_{i=0}^n$ , alors il existe  $\xi \in I$  tel que l'erreur d'interpolation au point  $x \in I$  est donnée par

$$E_{n-1}(x) \stackrel{\text{def}}{=} f(x) - p_{n-1}(x) = \frac{f^{(n)}(\xi)}{n!} \omega_n(x)$$

avec  $p_{n-1} \in \mathbb{R}_{n-1}[x]$  le polynôme d'interpolation.

Dans le cas d'une distribution uniforme de noeuds, i.e. quand  $x_i = x_{i-1} + h$  avec  $i = 1, 2, \dots, n$  et  $h > 0$  et  $x_0$  donnés, on a

$$|\omega_n(x)| \leq (n-1)! \frac{h^n}{4}$$



(a) Distribution équairepartie des nœuds

(b) Nœuds de CHEBYSHEV-GAUSS-LOBATTO

FIGURE 2.1 – Interpolation de LAGRANGE, exemple de RUNGE

et donc

$$\max_{x \in I} |E_{n-1}(x)| \leq \frac{\max_{x \in I} |f^{(n)}(x)|}{4n} h^n.$$

Malheureusement, on ne peut pas déduire de cette relation que l'erreur tend vers 0 quand  $n$  tend vers l'infini, bien que  $h^n/[4n]$  tend effectivement vers 0. En fait, il existe des fonctions  $f$  pour lesquelles  $\max_{x \in I} |E_{n-1}(x)| \xrightarrow[n \rightarrow +\infty]{} +\infty$ . Ce résultat frappant indique qu'en augmentant le degré  $n$  du polynôme d'interpolation, on n'obtient pas nécessairement une meilleure reconstruction de  $f$ .

Ce phénomène est bien illustré par la fonction de RUNGE : soit la fonction  $f: [-5, 5] \rightarrow \mathbb{R}$  définie par  $f(x) = \frac{1}{1+x^2}$ . La fonction  $f$  est infiniment dérivable sur  $[-5, 5]$  et  $|f^{(n)}(\pm 5)|$  devient très rapidement grand lorsque  $n$  tend vers l'infini. Si on considère une distribution uniforme des nœuds on voit que l'erreur tend vers l'infini quand  $n$  tend vers l'infini. Ceci est lié au fait que la quantité  $\max_{x \in [-5, 5]} |f^{(n)}(x)|$  tend plus vite vers l'infini que  $\frac{h^n}{4n}$  tend vers zéro. La figure 2.1a montre ses polynômes interpolants de degrés 3, 5 et 10 pour une distribution équairepartie des nœuds. Cette absence de convergence est également mise en évidence par les fortes oscillations observées sur le graphe du polynôme d'interpolation (absentes sur le graphe de  $f$ ), particulièrement au voisinage des extrémités de l'intervalle. Ce comportement est connu sous le nom de *phénomène de RUNGE*. On peut éviter le phénomène de RUNGE en choisissant correctement la distribution des nœuds d'interpolation. Sur un intervalle  $[a, b]$ , on peut par exemple considérer les nœuds de CHEBYSHEV-GAUSS-LOBATTO (voir figure 2.1b)

$$x_i = \frac{a+b}{2} - \frac{b-a}{2} \cos\left(\frac{\pi}{n-1}(i-1)\right), \quad \text{pour } i = 0, \dots, n$$

Pour cette distribution particulière de noeuds, il est possible de montrer que, si  $f$  est dérivable sur  $[a, b]$ , alors  $p_n$  converge vers  $f$  quand  $n \rightarrow +\infty$  pour tout  $x \in [a, b]$ . Les nœuds de CHEBYSHEV-GAUSS-LOBATTO, qui sont les abscisses des nœuds équirépartis sur le demi-cercle unité, se trouvent à l'intérieur de  $[a, b]$  et sont regroupés près des extrémités de l'intervalle. Écrire un script qui affiche les courbes des figures 2.1a et 2.1b en s'appuyant sur les fonction de l'exercice précédent.

### Correction

Ces figures peuvent être obtenues par les instructions suivantes :

```
f=@(x) [1./(1+x.^2)]; % La fonction de Runge

x = [-5:.1:5]; % Pour l'affichage on évaluera f et les polynomes en ces points
y = f(x);

% NOEUDS EQUIREPARTIS

% Construction des polynomes

% n=4 points => p in R_3[x]
x1 = [linspace(-5,5,4)];
y1 = f(x1);
y1interp = newton2([x1;y1]',x);
```

```
% n=8 points => p in R_7[x]
x2 = [linspace(-5,5,8)];
y2 = f(x2);
y2interp = newton2([x2;y2]',x);

% n=16 points => p in R_15[x]
x3 = [linspace(-5,5,16)];
y3 = f(x3);
y3interp = newton2([x3;y3]',x);

% Affichage
plot(x,f(x),'r-','LineWidth',2, ...
      x,y1interp,'b:','LineWidth',2, ...
      x,y2interp,'m-.','LineWidth',2, ...
      x,y3interp,'g--','LineWidth',2, ...
      x3,y3,'*','MarkerSize',10)
legend('f','p_3','p_7','p_{15}');
saveas(gcf, "runge_lagrange.png", "png");

% NOEUDS DE CHEBICHEF

% Construction des polynomes

% n=4 points => p in R_3[x]
x1 = -5*cos(pi*[0:3]/3);
y1 = f(x1);
y1interp = newton2([x1;y1]',x);

% n=8 points => p in R_7[x]
x2 = -5*cos(pi*[0:7]/7);
y2 = f(x2);
y2interp = newton2([x2;y2]',x);

% n=16 points => p in R_15[x]
x3 = -5*cos(pi*[0:15]/15);
y3 = f(x3);
y3interp = newton2([x3;y3]',x);

% Affichage
plot(x,f(x),'r-','LineWidth',2, ...
      x,y1interp,'b:','LineWidth',2, ...
      x,y2interp,'m-.','LineWidth',2, ...
      x,y3interp,'g--','LineWidth',2, ...
      x3,y3,'*','MarkerSize',10)
legend('f','p_3','p_7','p_{15}');
saveas(gcf, "runge_lagrangeTGL.png", "png");
```

### ★ Exercice 2.13

Calculer numériquement le polynôme caractéristique  $p_{\mathbb{A}}(\lambda) \stackrel{\text{def}}{=} \det(\mathbb{A} - \lambda \mathbb{I})$  de la matrice suivante

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

### Correction

Il s'agit d'une matrice d'ordre 3 donc son polynôme caractéristique appartient à  $\mathbb{R}_3[\lambda]$  :

$$p_{\mathbb{A}}(\lambda) = \alpha_0 + \alpha_1 \lambda + \alpha_2 \lambda^2 + \alpha_3 \lambda^3.$$

Si on connaît la valeur du polynôme caractéristique en 4 points, par exemple

$\lambda_i$	-1	0	1	2
$p_{\mathbb{A}}(\lambda_i)$	$\det(\mathbb{A} + \mathbb{I})$	$\det(\mathbb{A})$	$\det(\mathbb{A} - \mathbb{I})$	$\det(\mathbb{A} - 2\mathbb{I})$

on pourra le déterminer de façon unique par interpolation :

```
clear all
A=[2 1 0; 1 2 1; 0 1 2]
Id=eye(size(A));
% Point d interpolation
Px=[-1 0 1 2];
% Valeur du polynome en les points d interpolation
for i=1:length(Px)
    Py(i)=det(A-Px(i)*Id);
end
% Calcul des coefficients du polynome
P=[Px',Py']
alphaInt=naivePoly(P)
```

Donc

$$p_{\mathbb{A}}(\lambda) = 4 - 10\lambda + 6\lambda^2 - \lambda^3.$$

On peut afficher l'allure du polynôme comme suit :

```
x=[-2:0.1:3];
yinterpol=naiveEval(alphaInt,x);
plot(x,yinterpol,'LineWidth',2, P(:,1),P(:,2), 'o', 'MarkerSize',10);
```

### Exercice 2.14

Soit  $f$  une fonction continue dont on connaît les valeurs uniquement pour  $t$  entier, c'est-à-dire on suppose connues les valeurs  $f(\kappa)$  pour tout  $\kappa \in \mathbb{Z}$ . Si  $t \in \mathbb{R} \setminus \mathbb{Z}$ , on définit une approximation  $p(t)$  de  $f(t)$  en interpolant la fonction  $f$  par un polynôme de degré 3 aux quatre points entiers les plus proches de  $t$ . Calculer  $p(t)$  et écrire un algorithme qui fournit  $p(t)$ .

#### Correction

Soit  $\ell = E[t]$  la partie entière<sup>3</sup> de  $t$ . Alors  $t \in [\ell; \ell + 1]$  et il s'agit de définir le polynôme  $p$  interpolant les points

$$(\kappa - 1, f(\kappa - 1)), \quad (\kappa, f(\kappa)), \quad (\kappa + 1, f(\kappa + 1)), \quad (\kappa + 2, f(\kappa + 2)),$$

ce qui donne

$$\begin{aligned} P(t) &= \sum_{i=0}^3 \left( f(\kappa - 1 + i) \prod_{\substack{j=0 \\ j \neq i}}^3 \frac{t - (\kappa - 1 + j)}{(\kappa - 1 + i) - (\kappa - 1 + j)} \right) = \sum_{i=0}^3 \left( f(\kappa - 1 + i) \prod_{\substack{j=0 \\ j \neq i}}^3 \frac{t - \kappa + 1 - j}{i - j} \right) \\ &= -\frac{f(\kappa - 1)}{6} (t - \kappa)(t - \kappa - 1)(t - \kappa - 2) + \frac{f(\kappa)}{2} (t - \kappa + 1)(t - \kappa - 1)(t - \kappa - 2) \\ &\quad - \frac{f(\kappa + 1)}{2} (t - \kappa + 1)(t - \kappa)(t - \kappa - 2) + \frac{f(\kappa + 2)}{6} (t - \kappa + 1)(t - \kappa)(t - \kappa - 1) \end{aligned}$$

**Require:**  $f: \mathbb{Z} \rightarrow \mathbb{R}$ ,  $t$

```

 $\kappa \leftarrow E[t]$ 
 $x_0 \leftarrow \kappa - 1$ 
 $x_1 \leftarrow \kappa$ 
 $x_2 \leftarrow \kappa + 1$ 
 $x_3 \leftarrow \kappa + 2$ 
 $y \leftarrow 0$ 
for  $i = 0$  to  $3$  do
     $L \leftarrow 1$ 
```

3. Pour tout nombre réel  $x$ , la partie entière notée  $E(x)$  est le plus grand entier relatif inférieur ou égal à  $x$ . Par exemple,  $E(2.3) = 2$ ,  $E(-2) = -2$  et  $E(-2.3) = -3$ . La fonction partie entière est aussi notée  $[x]$  (ou  $\lfloor x \rfloor$  par les anglo-saxons). On a toujours  $E(x) \leq x < E(x) + 1$  avec égalité si et seulement si  $x$  est un entier relatif. Pour tout entier relatif  $k$  et pour tout nombre réel  $x$ , on a  $E(x + k) = E(x) + k$ . L'arrondi à l'entier le plus proche d'un réel  $x$  peut être exprimé par  $E(x + 0.5)$ .

```

for  $j = 0$  to  $3$  do
  if  $j \neq i$  then
     $L \leftarrow \frac{t - x_j}{x_i - x_j} \times L$ 
  end if
end for
 $y \leftarrow y + f(x_i) \times L$ 
end for
return  $y$ 

```

**Exercice 2.15**

Pour calculer le zéro d'une fonction  $y = f(x)$  inversible sur un intervalle  $[a; b]$  on peut utiliser l'interpolation : après avoir évalué  $f$  sur une discréétisation  $x_i$  de  $[a; b]$ , on interpole l'ensemble  $\{(y_i, x_i)\}_{i=0}^n$  et on obtient un polynôme  $x = p(y)$  tel que

$$f(x) = 0 \iff x = p(0).$$

Utiliser cette méthode pour évaluer l'unique racine  $\alpha$  de la fonction  $f(x) = e^x - 2$  dans l'intervalle  $[0; 1]$  avec trois points d'interpolation.

**Correction**

Calculons d'abord les valeurs à interpoler

$i$	$x_i$	$y_i$
0	0	-1
1	$\frac{1}{2}$	$\sqrt{e} - 2$
2	1	$e - 2$

Le polynôme d'interpolation de LAGRANGE de degré  $n$  sur l'ensemble des  $n+1$  points  $\{(y_i, x_i)\}_{i=0}^n$  s'écrit

$$p_n(y) = \sum_{i=0}^n \left( x_i \prod_{\substack{j=0 \\ j \neq i}}^n \frac{y - y_j}{y_i - y_j} \right).$$

Ici  $n = 2$  donc on a

$$\begin{aligned} p(y) &= x_0 \frac{(y - y_1)(y - y_2)}{(y_0 - y_1)(y_0 - y_2)} + x_1 \frac{(y - y_0)(y - y_2)}{(y_1 - y_0)(y_1 - y_2)} + x_2 \frac{(y - y_0)(y - y_1)}{(y_2 - y_0)(y_2 - y_1)} \\ &= \frac{1}{2} \frac{(y+1)(y-e+2)}{(\sqrt{e}-2+1)(\sqrt{e}-2-e+2)} + \frac{(y+1)(y-\sqrt{e}+2)}{(e-2+1)(e-2-\sqrt{e}+2)}. \end{aligned}$$

Par conséquent une approximation de la racine de  $f$  est  $p(0) = \frac{1}{2} \frac{-e+2}{(\sqrt{e}-2+1)(\sqrt{e}-2-e+2)} + \frac{-\sqrt{e}+2}{(e-2+1)(e-2-\sqrt{e}+2)} \approx 0.7087486785$ .

*Remarque :* comme il n'y a que trois points d'interpolation, on pourrait calculer directement le polynôme interpolateur de  $f$  plutôt que de sa fonction réciproque et chercher les zéros de ce polynôme directement car il s'agit d'un polynôme de degré 2. Cependant cette idée ne peut pas être généralisée au cas de plus de trois points d'interpolation car on ne connaît pas de formule générale pour le calcul des zéros d'un polynôme de degré  $n \geq 3$ .

**Interpolation trigonométrique****Exercice 2.16**

Considérons la fonction  $f: [0; 2\pi] \rightarrow \mathbb{R}$  définie par  $f(x) = x(x - 2\pi)e^{-x}$ . Calculer  $\tilde{f}(x)$  lorsque  $n = 9$  et comparer graphiquement les fonctions  $f$  et  $\tilde{f}$ .

**Correction**

On commence par définir la fonction  $f$  et calculer les valeurs de  $f$  aux nœuds  $x_j = j\pi/5$ ,  $j = 0, \dots, 9$  à l'aide des instructions suivantes

```
% fonction à interpoler
f=@(x) [x.*(x-2*pi).*exp(-x)];
```

```
% points d'interpolation
n=9; % n+1 points
Px=2*pi/(n+1)*[0:n];
Py=f(Px);
```

On calcule alors le vecteur des coefficients de FOURIER :

```
if rem(n,2)==0
    M=n/2;
    mu=0;
else
    M=(n-1)/2;
    mu=1;
end

% Il faut un shift car on ne peut pas utiliser des indices negatifs ou nuls
for k=-M-mu:M
    c(k+M+mu+1)=1/(n+1)*sum(f(Px).*exp(-i*k*Px));
end
```

On peut comparer notre calcul avec celui effectué par Octave grâce à la FFT et vérifier que la norme de l'erreur est nulle :

```
C=fftshift(fft(f(Px)))/(n+1);
norm(c-C)
```

La valeur de  $\tilde{f}$  en un point  $x$  est égale à

```
sum(c.*exp(i*[-M-mu:M]*x))
```

Pour comparer graphiquement  $f$  et son interpolée, on doit calculer  $f$  et  $\tilde{f}$  sur  $[0;2\pi]$  en  $S$  points :

```
S=100;
x=2*pi/S*[0:S-1];

% valeur exacte
fx=f(x);

% tilde_f calcule par Octave
z = interpft (Py ,S);

% tilde_f calcule par notre formule
for s=1:S
    ftildex(s)=sum(c.*exp(i*[-M-mu:M]*x(s)));
end

plot(x,fx,'LineWidth',2,'r-',...
      x,ftildex,'LineWidth',2,'b:',...
      x,z,'LineWidth',2,'m.',...
      Px,Py,'o');
```

TO DO : il est intéressant de vérifier numériquement qu'en augmentant  $n$  l'approximation  $\tilde{f}$  converge vers  $f$  tandis qu'avec l'interpolation polynomiale ceci n'est pas vrai.

# Chapitre 3

## Approximation au sens des moindres carrés

Nous avons déjà vu que si  $n$  est grand, le polynôme d'interpolation de  $\mathbb{R}_n[x]$  n'est pas toujours une bonne approximation d'une fonction donnée/cherchée. De plus, si les données sont affectées par des erreurs de mesure, l'interpolation peut être instable. Ce problème peut être résolu avec l'interpolation composite (avec des fonctions linéaires par morceau ou des splines). Néanmoins, aucune de ces méthodes n'est adaptée à l'extrapolation d'informations à partir des données disponibles, c'est-à-dire, à la génération de nouvelles valeurs en des points situés à l'extérieur de l'intervalle contenant les noeuds d'interpolation. On introduit alors la méthode des moindres carrés : soit  $d_i = y_i - f(x_i)$  l'écart vertical du point  $(x_i, y_i)$  par rapport à la fonction  $f$ . La méthode des moindres carrés est celle qui choisit  $f$  de sorte que la somme des carrés de ces déviations soit minimale. Étant donné que la fonction  $\sum_i d_i$  est en générale une fonction de plusieurs variables, nous allons rappeler comment on minimise une telle fonction.

### 3.1 Optimisation de fonctions de plusieurs variables

- ★ Une fonction  $f$  de  $\mathbb{R}^n$  et à valeurs réelles fait correspondre à tout point  $\mathbf{x} \equiv (x_1, x_2, \dots, x_n)$  de  $\mathbb{R}^n$  au plus un réel  $f(\mathbf{x})$ .  $\mathbf{x} \in \mathbb{R}^n$  se note aussi  $\mathbf{x}$  ou  $\underline{x}$ . Si  $n = 2$ , on utilise souvent la notation  $(x, y)$ , si  $n = 3$  la notation  $(x, y, z)$ .
- ★ Le domaine de définition de  $f$  est l'ensemble  $\mathcal{D}_f \subset \mathbb{R}^n$  des points  $\mathbf{x} \equiv (x_1, x_2, \dots, x_n)$  qui ont une image par  $f$ .
- ★ L'image par  $f$  de  $\mathcal{D}$  est l'ensemble  $\text{Im}_f(\mathcal{D}_f) = \{r \in \mathbb{R} \mid r = f(\mathbf{x}), \mathbf{x} \in \mathbb{R}^n\} \subset \mathbb{R}$ .
- ★ L'ensemble des points  $S = \{(\mathbf{x}, f(\mathbf{x})) \mid \mathbf{x} \in \mathcal{D}_f\}$  de  $\mathbb{R}^{n+1}$  est la surface représentative de  $f$  ; c'est l'analogie de la courbe représentative d'une fonction d'une variable. Évidemment, la représentation géométrique devient plus lourde que pour les fonctions d'une seule variable : une fonction de  $n$  variables se visualise à priori dans un espace à  $n + 1$  dimensions ( $n$  pour les variables, 1 pour le résultat de la fonction), alors que les pages d'un livre sont, par nature, bidimensionnelles. Pour contourner cette impossibilité technique, nous nous limiterons aux représentations des fonctions de deux variables, soit sous forme de dessins en perspective, soit sous forme de coupes par des plans horizontaux ou verticaux qui donnent des informations souvent utiles, quoique parcellaires. Ce problème de visualisation introduit une rupture nette par rapport aux fonctions d'une variable étudiées antérieurement.

Lorsque  $n = 2$ , le graphe

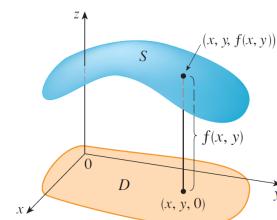
$$\mathcal{G}_f \equiv \{(x, y, z = f(x, y)) \mid (x, y) \in \mathcal{D}_f\}$$

est tridimensionnel. On peut considérer le graphe d'une fonction de deux variables comme étant le relief d'une région (par exemple, l'altitude en fonction de la longitude et de la latitude).

On visualise le graphe d'une fonction

$$\begin{aligned} f: \mathbb{R}^2 &\rightarrow \mathbb{R} \\ (x, y) &\mapsto f(x, y) \end{aligned}$$

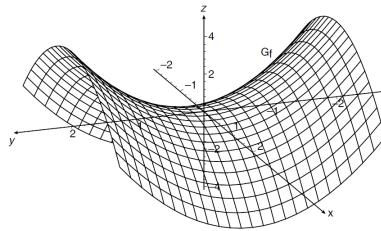
par l'altitude  $z = f(x, y)$ .



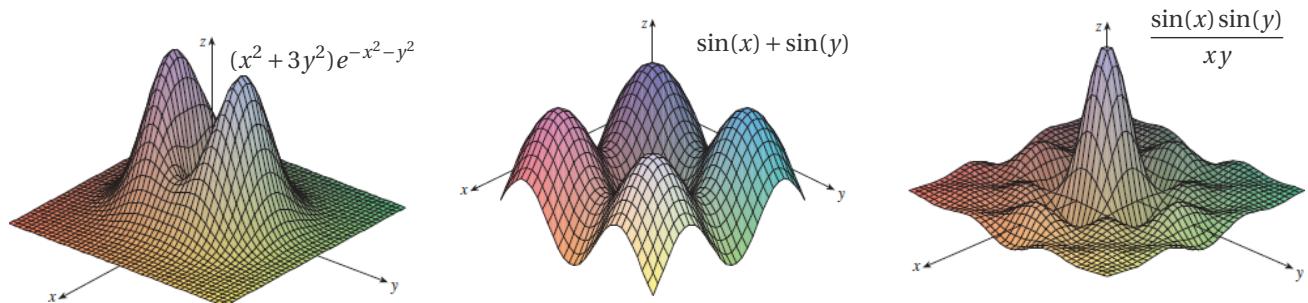
Les axes relatifs aux variables,  $x$  et  $y$ , sont conventionnellement situés dans un plan horizontal (le domaine  $\mathcal{D}_f$  apparaît alors comme un sous-ensemble de ce plan), tandis que la dimension verticale est réservée aux valeurs de  $z$ . Ainsi, à tout  $(x, y) \in \mathcal{D}_f$ , dont l'image est  $f(x, y) \in \mathbb{R}$ , correspond le point suivant du graphe :  $(x, y, f(x, y)) \in \mathbb{R}^3$ . Une mise en perspective permet la visualisation des surfaces à trois dimensions. Dans ce cas, l'axe  $z$  est toujours placé verticalement. Toutefois, pour des raisons de lisibilité, les axes  $x$  et  $y$  ne sont pas toujours présentés selon la même orientation.

#### EXEMPLE

Le graphe de la fonction  $f: \mathbb{R}^2 \rightarrow \mathbb{R}$  définie par  $f(x, y) = x^2 - y^2$  est une surface de  $\mathbb{R}^3$  qui a la forme d'une selle de cheval, comme l'indique la représentation en perspective de la figure ci-dessous.



Voici d'autres exemples :

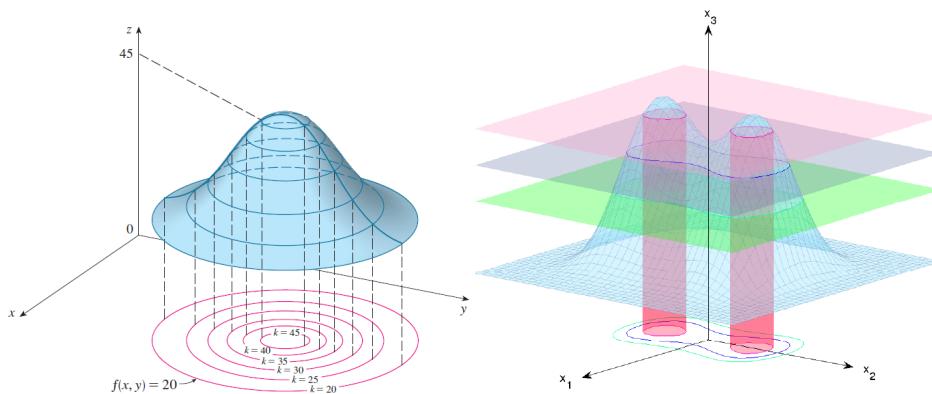


Si on considère des coupes horizontales on obtient, de façon générale, des courbes planes, dites *courbes ou lignes de niveau*.

### Définition 3.1 (Lignes de niveau)

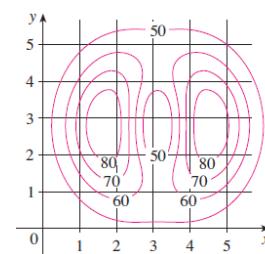
Soit  $k \in \mathbb{R}$  et  $f$  une fonction de  $\mathbb{R}^2$  dans  $\mathbb{R}$ ; la courbe de niveau  $k$  de la fonction  $f$  est la projection sur le plan d'équation  $z = 0$  de l'intersection de la surface représentative de  $f$  avec le plan horizontal  $z = k$ , i.e. l'ensemble  $\{(x, y) \in \mathcal{D} \mid f(x, y) = k\}$ .

En pratique, on représente simultanément différentes courbes de niveau pour visualiser la progression du graphe. Cette représentation s'apparente aux cartes géographiques où le niveau correspond à l'altitude. Les courbes de niveau d'une fonction  $f(x, y)$  fournissent une représentation géométrique de  $f$  sur le plan, alors que son graphe en donne une dans l'espace.



### EXEMPLE

L'image ci-contre montre les courbes de niveaux d'une fonction  $f$ . On peut alors se faire une idée de l'allure de la fonction supposée continue. Par exemple  $f(1; 3) \approx 72$ ,  $f(4; 5) \approx 56$ , soit  $40 < f(3; 3) < 50$  soit  $50 < f(3; 3) < 60$ , etc.

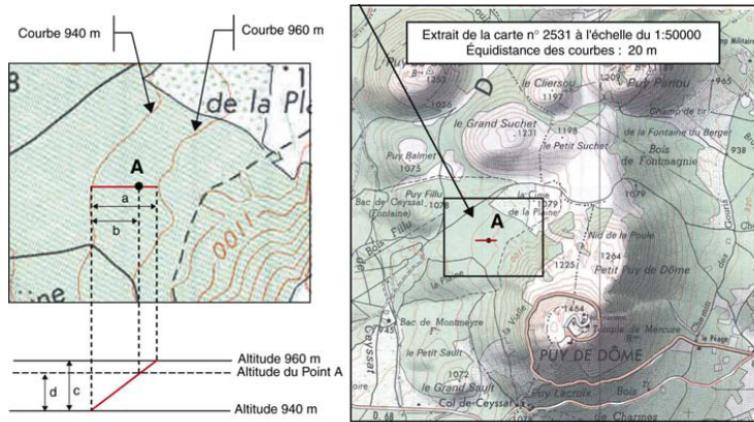


### EXEMPLE (CARTES TOPOGRAPHIQUES)

On peut considérer le relief d'une région comme étant le graphe d'une fonction de deux variables (par exemple, l'altitude en fonction de la longitude et de la latitude). Une courbe de niveau nous indique les points de même altitude (ici, l'altitude du point  $A$  est  $940 + d = 940 + cb/a$ ). En dessinant les courbes de niveau avec leur altitude correspondante, on obtient la *carte topographique du relief*. La lecture d'une carte topographique permet non seulement d'obtenir des mesures quantitatives

du relief, mais aussi de faire rapidement des observations qualitatives sur sa nature. Par exemple, localiser les points de plus haute et de plus basse altitude ; les crêtes, les fonds, les vallées, les cols, etc. ; les endroits du relief où les pentes sont plus escarpées ou plus douces, puisqu'ils correspondent respectivement aux courbes de niveau très rapprochées ou très distantes.

Attention : dans cette représentation les couleurs ne correspondent pas à la représentation planaire mais servent à reproduire les ombres.



### 3.1.1 Dérivées partielles du premier ordre et gradient

**Rappels** Soit  $f$  une fonction à valeurs réelles définie sur  $I$  un intervalle ouvert de  $\mathbb{R}$ . On dit que  $f$  est dérivable en  $x_0 \in I$  s'il existe finie la limite

$$\lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

ce qui équivaut, en posant  $h = x - x_0$ , à

$$\lim_{h \rightarrow 0} \frac{f(x_0 + h) - f(x_0)}{h}.$$

Cette limite est notée  $f'(x_0)$  et appelée dérivée de  $f$  en  $x_0$ .

L'unique dérivée d'une fonction d'une variable réelle, lorsqu'elle existe, est liée aux variations de la fonction tandis que la variable parcourt l'axe des abscisses. Pour une fonction  $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ , dont le graphe est une surface de  $\mathbb{R}^3$ , la situation est très différente. En effet, l'axe réel n'offre que deux types de mouvements possibles : de gauche à droite et de droite à gauche tandis que le plan  $\mathbb{R}^2$  possède une infinité de directions. Il peut s'avérer intéressant d'étudier comment une fonction  $f: \mathbb{R}^2 \rightarrow \mathbb{R}$  évolue lorsque la variable suit l'une ou l'autre direction du plan. À cet égard, considérons d'abord la direction à  $y$  fixé. Prenons le point  $(x_0, y_0)$  du domaine de  $f$ . Son image est  $f(x_0, y_0) \in \mathbb{R}$  et le graphe de la fonction, qui est la surface d'équation  $z = f(x, y)$  de  $\mathbb{R}^3$ , comporte le point  $(x_0, y_0, f(x_0, y_0))$ . L'intersection du graphe de  $f$  avec le plan vertical  $y = y_0$  est la courbe d'équation  $z = f(x, y_0)$  de  $\mathbb{R}^2$ . Le point  $(x_0, y_0)$  étant fixé, on peut alors interpréter cette courbe comme le graphe de la fonction  $f_{y=y_0}$  d'une seule variable définie par  $f_{y=y_0}(x) = f(x, y_0)$  dans le repère  $x0z$ . Si  $f_{y=y_0}$  est dérivable en  $x_0$ , alors sa dérivée nous renseigne sur la variation de la fonction  $f$  lorsque  $(x, y)$  se déplace le long de la droite horizontale de  $\mathbb{R}^2$  passant par le point  $(x_0, y_0)$ . Par analogie on peut répéter le même raisonnement à  $x$  fixé. En conclusion, lorsqu'on pose toutes les variables d'une fonction égales à une constante, sauf une, on obtient alors une fonction d'une seule variable qui peut être dérivée suivant les règles habituelles.



### **Définition 3.2 (Dérivées partielles premières)**

Soit  $f$  une fonction à valeurs réelles définie sur une partie ouverte  $\mathcal{D}$  de  $\mathbb{R}^2$ . Soit  $(x_0, y_0) \in \mathcal{D}$ . Les dérivées partielles de  $f$  en  $(x_0, y_0)$  sont les dérivées des fonctions partielles  $f_{y_0}$  et  $f_{x_0}$  évaluées en  $(x_0, y_0)$ , i.e. les fonctions

$$\frac{\partial f}{\partial x}(x_0, y_0) = \lim_{x \rightarrow x_0} \frac{f(x, y_0) - f(x_0, y_0)}{x - x_0} = \lim_{h \rightarrow 0} \frac{f(x_0 + h, y_0) - f(x_0, y_0)}{h}$$

$$\frac{\partial f}{\partial y}(x_0, y_0) = \lim_{k \rightarrow 0} \frac{f(x_0, y) - f(x_0, y_0)}{y - y_0} = \lim_{k \rightarrow 0} \frac{f(x_0, y_0 + k) - f(x_0, y_0)}{k}$$

dérivée partielle de  $f$   
par rapport à  $x$  au point  $(x_0, y_0)$

dérivée partielle de  $f$   
par rapport à  $y$  au point  $(x_0, y_0)$

*Il s'agit de limites d'une fonction réelle de variable réelle !*

Si  $f$  admet toutes les dérivées partielles premières, on dit que  $f$  est *dérivable*.

### ✿ Remarque (Notation)

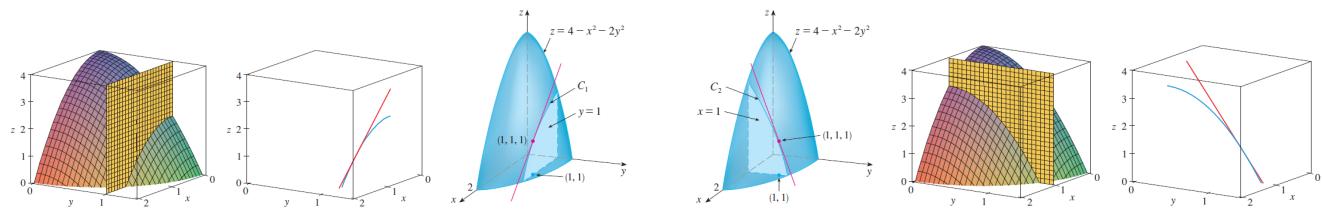
La dérivée  $\frac{\partial f}{\partial x}$  se note aussi  $\partial_x f$  ou  $f_{,x}$  ou encore  $\left. \frac{\partial f}{\partial x} \right|_y$  en insistant sur la variable qu'on considère constante.  
(Attention à ne pas confondre  $f_{,x}$  la dérivée de  $f$  par rapport à  $x$  avec  $f_{x=x_0}$  la fonction partielle associée à  $f$ .)

### ☛ Astuce

En pratique, pour calculer la dérivée partielle  $\partial_x f$  (resp.  $\partial_y f$ ), on dérive  $f$  comme si elle était une fonction de la seule variable  $x$  (resp.  $y$ ) et que l'autre variable,  $y$  (resp.  $x$ ), était une constante.

### ☛ EXEMPLE

Soit  $f(x, y) = 4 - x^2 - 2y^2$ . Le graphe de  $f$  est le paraboloïde  $z = 4 - x^2 - 2y^2$ . On a  $\partial_x f(x, y) = -2x$  et  $\partial_y f(x, y) = -4y$ .  
Le plan vertical  $y = 1$  intersecte le paraboloïde dans la parabole d'équation  $z(x) = 2 - x^2$  (et on appelle cette courbe  $C_1$  comme dans la figure à gauche). La pente de la droite tangente à cette parabole au point  $(1, 1)$  est  $\partial_x f(1, 1) = -2$ .  
De la même façon, le plan vertical  $x = 1$  intersecte le paraboloïde dans la parabole  $z(y) = 2 - 2y^2$  (et on appelle cette courbe  $C_2$  comme dans la figure à droite). La pente de la droite tangente à cette parabole au point  $(1, 1)$  est  $\partial_y f(1, 1) = -4$ .



### ☛ EXEMPLE

1. Soit la fonction  $f(x, y) = 3x^2 + xy - 2y^2$ . Alors  $\mathcal{D}_f \equiv \mathbb{R}^2$ ,  $f$  est continue,  $\partial_x f(x, y) = 6x + y$  (car  $y$  est considérée constante) et  $\partial_y f(x, y) = x - 4y$  (car  $x$  est considérée constante).
2. Soit la fonction  $f(x, y, z) = 5xz \ln(1 + 7y)$ . Alors  $\mathcal{D}_f \equiv \{(x, y, z) \mid y > -1/7\}$ ,  $f$  est continue et  $\partial_x f(x, y, z) = 5z \ln(1 + 7y)$ ,  $\partial_y f(x, y, z) = \frac{35xz}{1+7y}$  et  $\partial_z f(x, y, z) = 5x \ln(1 + 7y)$ .
3. La résistance totale  $R$  d'un conducteur produite par trois conducteurs de résistances  $R_1, R_2, R_3$ , connectés en parallèle, est donnée par la formule

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}.$$

On a alors  $\partial_{R_i} R(R_1, R_2, R_3) = R^2 / R_i^2$ .



### Définition 3.3 (Vecteur gradient)

Le gradient de la fonction  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  évalué au point  $\hat{\mathbf{x}} = (\hat{x}_1, \hat{x}_2, \dots, \hat{x}_n)$ , noté  $\nabla f(\hat{\mathbf{x}})$  ou encore  $\text{grad } f(\hat{\mathbf{x}})$ , est le vecteur dont les composantes sont les dérivées partielles premières :

$$\nabla f(\hat{\mathbf{x}}) = \begin{pmatrix} \partial_{x_1} f(\hat{\mathbf{x}}) \\ \partial_{x_2} f(\hat{\mathbf{x}}) \\ \vdots \\ \partial_{x_n} f(\hat{\mathbf{x}}) \end{pmatrix}$$

Il est **orthogonal** à la courbe de niveau de  $f$  passant par  $\hat{\mathbf{x}}$ .



### Définition 3.4 (Plan tangent)

Soit  $\mathcal{D}$  une partie ouverte de  $\mathbb{R}^n$  et soit  $f: \mathcal{D} \rightarrow \mathbb{R}$  une fonction différentiable en  $\hat{\mathbf{x}}$ . L'équation du plan tangent au graphe de la fonction  $f(\mathbf{x})$  en  $\hat{\mathbf{x}}$  est

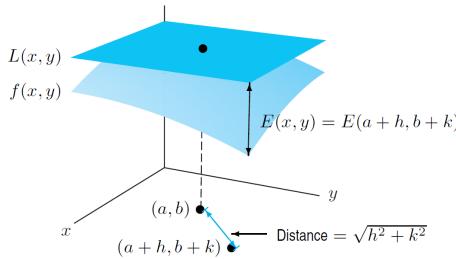
$$L(\mathbf{x}) = f(\hat{\mathbf{x}}) + (\mathbf{x} - \hat{\mathbf{x}})^T \cdot \nabla f(\hat{\mathbf{x}}).$$

Pour  $n = 2$ , en notant  $\hat{\mathbf{x}} \equiv (x_0, y_0)$ , l'équation du plan tangent au graphe de la fonction  $f(x, y)$  en  $(x_0, y_0)$  s'écrit

$$L(x, y) = f(x_0, y_0) + (x - x_0)\partial_x f(x_0, y_0) + (y - y_0)\partial_y f(x_0, y_0)$$

ce qui équivaut, en notant  $(h, k) = (x - x_0, y - y_0)$ , à

$$L(x_0 + h, y_0 + k) = f(x_0, y_0) + h\partial_x f(x_0, y_0) + k\partial_y f(x_0, y_0).$$



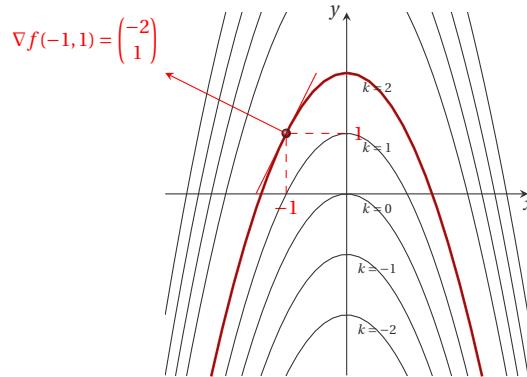
Graphe de la fonction  $(x, y) \mapsto f(x, y)$  et sa linéarisation  $(x, y) \mapsto L(x, y)$  au voisinage du point  $(a, b)$ . La fonction  $(h, k) \mapsto E(h, k) \equiv f(a+h, b+k) - L(a+h, b+k)$  mesure l'erreur qu'on fait au point  $(a+h, b+k)$  lorsqu'on approche la valeur de  $f$  par la valeur du plan tangent  $L$ . Si  $f$  est différentiable au point  $(a, b)$  alors

$$\lim_{(h,k) \rightarrow (0,0)} \frac{E(h, k)}{\sqrt{h^2 + k^2}} = 0.$$

### EXEMPLE

Considérons la fonction de  $\mathbb{R}^2$  dans  $\mathbb{R}$  définie par  $f(x, y) = x^2 + y$ . Le gradient de  $f$  est le vecteur  $\nabla f(x, y) = (2x, 1)^T$ . La courbe de niveau  $k$  de la fonction  $f$  est l'ensemble  $\{(x, y) \in \mathbb{R}^2 \mid x^2 + y = k\}$ , autrement dit la parabole d'équation  $y = -x^2 + k$ . Le gradient est orthogonal à la courbe de niveau de  $f$  qui passe par le point  $(x, y)$ .

Dans la figure ci-dessous on considère le point  $(-1, 1)$ . Le vecteur gradient de  $f$  dans ce point vaut  $(-2, 1)^T$ . Le point donné appartient à la courbe de niveau 2 qui a pour équation  $y = -x^2 + 2$ . La droite tangente à cette courbe au point  $(-1, 1)$  a pour équation  $y = 2x + 3$  qui est orthogonale au gradient.



Le plan tangent à  $f$  en  $\hat{x} = (-1, 1)$  s'écrit

$$L(\mathbf{x}) = f(\hat{\mathbf{x}}) + (\mathbf{x} - \hat{\mathbf{x}})^T \cdot \nabla f(\hat{\mathbf{x}}) = (-1)^2 + 1 + (x+1, y-1) \cdot \begin{pmatrix} -2 \\ 1 \end{pmatrix} = 2 - 2(x+1) + (y-1) = -2x + y - 1.$$

Cette notion se généralise naturellement pour  $n > 2$  : il s'agit en fait d'un plan tangent pour  $n = 2$  et d'un hyperplan tangent pour  $n > 2$ . Dans un espace de dimension  $n$ , un hyperplan est une variété linéaire de dimension  $n - 1$ .

### EXEMPLE

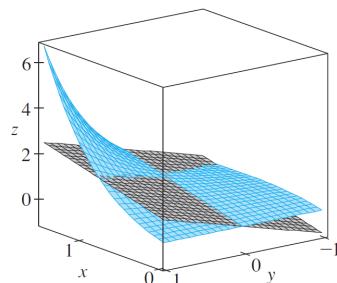
On peut calculer le plan tangent à la fonction  $f: \mathbb{R}^2 \rightarrow \mathbb{R}$  définie par  $f(x, y) = xe^{xy}$  en  $(1, 0)$  et utiliser sa linéarisation pour approcher  $f(1.1, -0.1)$ . On a

$$\begin{aligned} f(x, y) &= xe^{xy} & f(1, 0) &= 1 \\ \partial_x f(x, y) &= e^{xy} + xy e^{xy} & \partial_x f(1, 0) &= 1 \\ \partial_y f(x, y) &= x^2 e^{xy} & \partial_y f(1, 0) &= 1 \end{aligned}$$

Les trois fonctions  $f$ ,  $\partial_x f$  et  $\partial_y f$  sont continues, donc  $f$  est différentiable. Sa linéarisation donne

$$f(x, y) \approx f(1, 0) + (x-1)\partial_x f(1, 0) + (y-0)\partial_y f(1, 0) = 1 + (x-1) + y = x + y,$$

autrement dit  $xe^{xy} \approx x + y$  lorsque  $(x, y) \approx (1, 0)$ , ainsi  $f(1.1, -0.1) \approx 1.1 - 0.1 = 1$ . En effet,  $f(1.1, -0.1) = 1.1e^{-0.11} \approx 0.98542$



### 3.1.2 Dérivées partielles de deuxième ordre et matrice hessienne

Si les fonctions dérivées partielles admettent elles-mêmes des dérivées partielles en  $(x_0, y_0)$ , ces dérivées sont appelées dérivées partielles secondes, ou dérivées partielles d'ordre 2, de  $f$  en  $(x_0, y_0)$ . On peut, de la même façon, introduire les dérivées partielles d'ordres supérieurs. Les définitions suivantes s'énoncent dans des ensembles ouverts pour éviter les problèmes liés au calcul de limites au bord du domaine.



#### Définition 3.5 (Dérivées partielles d'ordre 2 pour une fonction de deux variables)

Soit la fonction  $f: \mathcal{D} \subset \mathbb{R}^2 \rightarrow \mathbb{R}$  où  $\mathcal{D}_f$  est un ouvert de  $\mathbb{R}^2$ . On a 2 dérivées partielles d'ordre 1 et donc 4 dérivées partielles d'ordre 2 ainsi notées :

$$\begin{aligned}\frac{\partial^2 f}{\partial x^2}(x_0, y_0) &= \frac{\partial}{\partial x} \left( \frac{\partial f}{\partial x} \right)(x_0, y_0) && \text{(notée aussi } \partial_{xx} f(x_0, y_0)), \\ \frac{\partial^2 f}{\partial x \partial y}(x_0, y_0) &= \frac{\partial}{\partial x} \left( \frac{\partial f}{\partial y} \right)(x_0, y_0) && \text{(notée aussi } \partial_{xy} f(x_0, y_0)), \\ \frac{\partial^2 f}{\partial y \partial x}(x_0, y_0) &= \frac{\partial}{\partial y} \left( \frac{\partial f}{\partial x} \right)(x_0, y_0) && \text{(notée aussi } \partial_{yx} f(x_0, y_0)), \\ \frac{\partial^2 f}{\partial y^2}(x_0, y_0) &= \frac{\partial}{\partial y} \left( \frac{\partial f}{\partial y} \right)(x_0, y_0) && \text{(notée aussi } \partial_{yy} f(x_0, y_0)).\end{aligned}$$

Les dérivées partielles d'ordre supérieur à 2 se définissent par récurrence de façon analogue. Soit la fonction  $f: \mathbb{R}^n \rightarrow \mathbb{R}$ ; on aura  $n$  dérivées partielles d'ordre 1,  $n^2$  dérivées partielles d'ordre 2, etc. donc  $n^k$  dérivées partielles d'ordre  $k$ .



#### Théorème 3.6 (Théorème de SCHWARZ (ou de CLAIRAUT))

Si les dérivées partielles mixtes  $\partial_{xy} f$  et  $\partial_{yx} f$  sont continues en  $(x_0, y_0)$  alors  $\partial_{xy} f(x_0, y_0) = \partial_{yx} f(x_0, y_0)$ .



#### Définition 3.7 (Matrice hessienne)

Soit la fonction  $f: \mathcal{D} \subset \mathbb{R}^2 \rightarrow \mathbb{R}$  où  $\mathcal{D}_f$  est un ouvert de  $\mathbb{R}^2$ . La matrice hessienne de  $f$  en  $(x_0, y_0)$  est la matrice de taille  $2 \times 2$  dont les entrées sont les dérivées partielles secondes :

$$H_f(x_0, y_0) = \begin{pmatrix} \partial_{xx} f(x_0, y_0) & \partial_{xy} f(x_0, y_0) \\ \partial_{yx} f(x_0, y_0) & \partial_{yy} f(x_0, y_0) \end{pmatrix}.$$

Son déterminant est le réel  $\det(H_f(x_0, y_0)) \equiv \partial_{xx} f(x_0, y_0) \partial_{yy} f(x_0, y_0) - \partial_{xy} f(x_0, y_0) \partial_{yx} f(x_0, y_0)$ .

Cette notion se généralise naturellement pour  $n > 2$ .



Les dérivées premières et secondes de la fonction  $f(x, y) = -2x^2 + 3xy^2 - y^3$  sont

$$\begin{aligned}\partial_x f(x, y) &= -4x + 3y^2, & \partial_y f(x, y) &= 6xy - 3y^2, \\ \partial_{xx} f(x, y) &= -4, & \partial_{xy} f(x, y) &= 6y, & \partial_{yx} f(x, y) &= 6y, & \partial_{yy} f(x, y) &= 6x - 6y.\end{aligned}$$

La matrice hessienne est

$$H_f(x, y) = \begin{pmatrix} -4 & 6y \\ 6y & 6x - 6y \end{pmatrix}.$$

Dans cet exemple, on remarque que la matrice hessienne de  $f$  est symétrique du fait que les dérivées secondes mixtes,  $\partial_{xy} f$  et  $\partial_{yx} f$ , sont égales.

Comme la dérivée seconde pour les fonctions d'une seule variable, la matrice hessienne permet d'étudier la convexité des fonctions de plusieurs variables et joue, dès lors, un rôle important dans leur optimisation.



#### Définition 3.8 (Convexité dans $\mathbb{R}^2$ )

Soit  $\mathcal{D}$  un sous-ensemble convexe de  $\mathbb{R}^2$  et  $f: \mathcal{D} \rightarrow \mathbb{R}$  une fonction.

- \*  $f$  est concave dans  $\mathcal{D}$  si

$$f((1-t)(x_0, y_0) + t(x_1, y_1)) \geq (1-t)f(x_0, y_0) + tf(x_1, y_1) \quad \forall (x_0, y_0), (x_1, y_1) \in \mathcal{D} \text{ et } \forall t \in [0; 1];$$

- \*  $f$  est strictement concave dans  $\mathcal{D}$  si

$$f((1-t)(x_0, y_0) + t(x_1, y_1)) > (1-t)f(x_0, y_0) + tf(x_1, y_1) \quad \forall (x_0, y_0), (x_1, y_1) \in \mathcal{D} \text{ et } \forall t \in ]0; 1[;$$

- ★  $f$  est convexe dans  $\mathcal{D}$  si

$$f((1-t)(x_0, y_0) + t(x_1, y_1)) \leq (1-t)f(x_0, y_0) + tf(x_1, y_1) \quad \forall (x_0, y_0), (x_1, y_1) \in \mathcal{D} \text{ et } \forall t \in [0; 1];$$

- ★  $f$  est strictement convexe dans  $\mathcal{D}$  si

$$f((1-t)(x_0, y_0) + t(x_1, y_1)) < (1-t)f(x_0, y_0) + tf(x_1, y_1) \quad \forall (x_0, y_0), (x_1, y_1) \in \mathcal{D} \text{ et } \forall t \in ]0; 1[;$$

Comme pour les fonctions d'une variable, la concavité et la convexité des fonctions de  $n$  variables suffisamment régulières peuvent être caractérisées à l'aide des dérivées d'ordres 1 ou 2.

### Propriété 3.9

Tout plan tangent au graphe d'une fonction concave (resp. convexe) se trouve au-dessus (resp. au-dessous) de ce graphe : soit  $\mathcal{D}$  un sous-ensemble convexe de  $\mathbb{R}^2$  et  $f: \mathcal{D} \rightarrow \mathbb{R}$  une fonction différentiable dans  $\mathcal{D}$ . Alors

1.  $f$  est concave dans  $\mathcal{D} \iff \forall (x_0, y_0), (x_1, y_1) \in \mathcal{D}, f(x_1, y_1) \leq f(x_0, y_0) + (x_1 - x_0)\partial_x f(x_0, y_0) + (y_1 - y_0)\partial_y f(x_0, y_0);$
2.  $f$  est convexe dans  $\mathcal{D} \iff \forall (x_0, y_0), (x_1, y_1) \in \mathcal{D}, f(x_1, y_1) \geq f(x_0, y_0) + (x_1 - x_0)\partial_x f(x_0, y_0) + (y_1 - y_0)\partial_y f(x_0, y_0).$

Si, de plus,  $f \in \mathcal{C}^2(\mathcal{D})$ , alors

1.  $\forall (x, y) \in \mathcal{D}, H_f(x, y)$  est définie négative (i.e.  $\det(H_f(x, y)) > 0$  et  $\partial_{xx} f(x, y) < 0$ )  $\implies f$  est strictement concave dans  $\mathcal{D}$ ;
2.  $\forall (x, y) \in \mathcal{D}, H_f(x, y)$  est définie positive (i.e.  $\det(H_f(x, y)) > 0$  et  $\partial_{xx} f(x, y) < 0$ )  $\implies f$  est strictement convexe dans  $\mathcal{D}$ .

Notons qu'une fonction peut être strictement convexe sans que sa matrice hessienne soit définie positive en tout point.

### EXEMPLE

1. Soit  $f(x, y) = x^2 + y^2$ . On a

- ★  $f \in \mathcal{C}^2(\mathbb{R}^2)$ ,
- ★  $H_f(x, y) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$  pour tout  $(x, y) \in \mathbb{R}^2$ ,
- ★  $\det(H_f(x, y)) = 4 > 0$  et  $\partial_{xx} f(x, y) = 2 > 0$  donc  $H_f(x, y)$  est définie positive pour tout  $(x, y) \in \mathbb{R}^2$ .

Il s'en suit que  $f$  est strictement convexe dans  $\mathbb{R}^2$ . En effet, pour tout  $(x, y) \in \mathbb{R}^2 \setminus \{(0, 0)\}$ ,

$$f(x, y) = x^2 + y^2 > f(0, 0) + x\partial_x f(0, 0) + y\partial_y f(0, 0).$$

2. Soit  $f(x, y) = x^4 + y^4$ . On a

- ★  $f \in \mathcal{C}^2(\mathbb{R}^2)$ ,
- ★  $H_f(x, y) = \begin{pmatrix} 12x^2 & 0 \\ 0 & 12y^2 \end{pmatrix}$  pour tout  $(x, y) \in \mathbb{R}^2$ ,
- ★  $\det(H_f(x, y)) = 144x^2y^2 \geq 0$  et  $\partial_{xx} f(x, y) = 12x^2 \geq 0$  donc  $H_f(x, y)$  est définie positive pour tout  $(x, y) \in (\mathbb{R}^*)^2$  et semi-définie positive en  $(0, 0)$ .

On peut cependant montrer à l'aide de la définition que la fonction est strictement convexe dans  $\mathbb{R}^2$  car

$$f(x, y) = x^4 + y^4 > f(0, 0) + x\partial_x f(0, 0) + y\partial_y f(0, 0) = 0 \quad \forall (x, y) \neq (0, 0).$$

### 3.1.3 Optimisation

Un optimum ou extremum est soit un maximum soit un minimum, c'est-à-dire la valeur la plus haute ou la plus faible que prend la fonction sur son ensemble de définition ou tout sous-ensemble de son ensemble de définition.

### Définition 3.10

Soit  $f$  une fonction de  $\mathcal{D} \subset \mathbb{R}^n$  dans  $\mathbb{R}$ . On dit que

- ★  $f$  est bornée dans  $\mathcal{D}$  s'il existe un nombre réel  $M \geq 0$  tel que

$$\forall \mathbf{x} \in \mathcal{D}, |f(\mathbf{x})| \leq M;$$

- ★  $f$  admet un maximum (resp. minimum) global (ou absolu) en  $\mathbf{x}_0 \in \mathcal{D}$  si

$$\forall \mathbf{x} \in \mathcal{D}, f(\mathbf{x}) \leq f(\mathbf{x}_0) \text{ (resp. } f(\mathbf{x}) \geq f(\mathbf{x}_0));$$

- \*  $f$  admet un maximum (resp. minimum) *local* (ou relatif) en  $\mathbf{x}_0 \in \mathcal{D}$  s'il existe une boule de rayon non nul  $\mathcal{B}(\mathbf{x}_0, r)$  telle que

$$\forall \mathbf{x} \in \mathcal{D} \cap \mathcal{B}(\mathbf{x}_0, r), f(\mathbf{x}) \leq f(\mathbf{x}_0) \text{ (resp. } f(\mathbf{x}) \geq f(\mathbf{x}_0)).$$



### Théorème 3.11 (de FERMAT : condition nécessaire du premier ordre)

Soit  $\mathcal{D}$  un sous-ensemble ouvert de  $\mathbb{R}^n$ ,  $\mathbf{x}_0$  un point contenu dans  $\mathcal{D}$  et  $f: \mathcal{D} \rightarrow \mathbb{R}$  une fonction de classe  $\mathcal{C}^1$  en ce point. Si  $f$  présente un extrémum local alors

$$\nabla f(\mathbf{x}_0) = \mathbf{0}.$$



### Définition 3.12 (Point stationnaire ou critique)

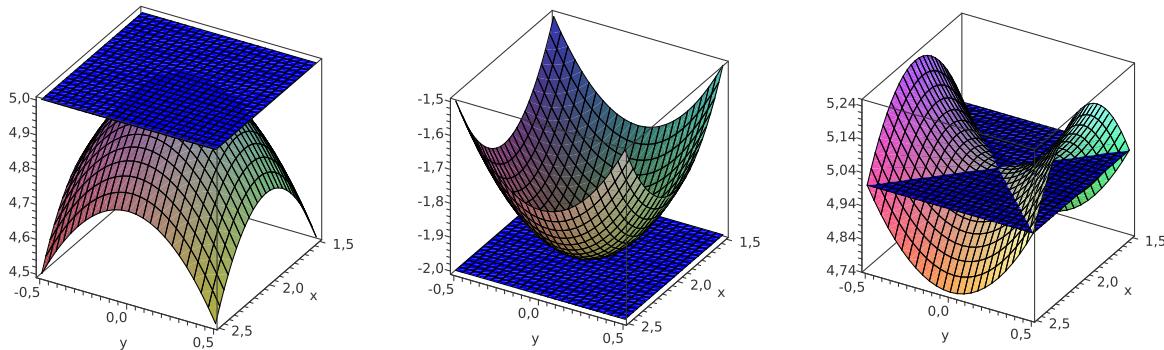
À l'instar des fonctions d'une variable réelle, un point  $\mathbf{x}_0$  vérifiant  $\nabla f(\mathbf{x}_0) = \mathbf{0}$  est appelé *point stationnaire* ou *point critique* de  $f$ .

**Nature d'un point critique : étude directe** La condition du premier ordre signifie géométriquement que le plan tangent à la surface d'équation  $z = f(x, y)$  au point  $(x_0, y_0)$  de coordonnées  $(x_0, y_0, f(x_0, y_0))$  est horizontal. Après avoir déterminé un point stationnaire  $\mathbf{x}_0$ , on peut alors déterminer sa nature en étudiant le signe de la différence

$$d(\mathbf{h}) = f(\mathbf{x}_0 + \mathbf{h}) - f(\mathbf{x}_0).$$

Si cette différence est de signe constant pour  $\mathbf{h}$  voisin de  $\mathbf{0}$ , il s'agit d'un extrémum local (un *maximum* si  $d < 0$ , un *minimum* si  $d > 0$ ). Sinon, il s'agit d'un *point-col* (ou *point-selle*). Mieux, si le signe est constant pour  $\mathbf{h}$  quelconque, alors l'extrémum est global.

La figure à gauche illustre le cas d'un *maximum* et la figure au centre le cas d'un *minimum*. La figure à droite illustre le fait que la condition nécessaire d'optimalité n'est pas une condition suffisante ; dans ce cas on dit que  $f$  présente un *col* en  $(x_0, y_0)$  ou que  $(x_0, y_0)$  est un *point-selle* de  $f$ . Le mot col vient de l'exemple de la fonction altitude et de la configuration (idéalisée) d'un col de montagne : minimum de la ligne de crête, maximum de la route, sans être un extrémum du paysage. Le mot selle vient de l'exemple d'une selle de cheval.



### EXEMPLE

On cherche les extrema de la fonction  $f(x, y) = x^2 + y^2$  dans le disque ouvert centré en  $(0, 0)$  de rayon 1, représenté par  $\mathcal{D} = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 < 1\}$ . Le seul candidat extrémum est l'unique point critique  $(0, 0)$  qu'on trouve en résolvant  $\partial_x f(x, y) = 0$  et  $\partial_y f(x, y) = 0$ . La définition implique de façon immédiate que  $f$  admet un minimum global en  $(0, 0)$ . En effet

$$f(x, y) = x^2 + y^2 \geq 0 = f(0, 0) \quad \forall (x, y) \in \mathcal{D}_f.$$

En revanche, la fonction n'admet aucun maximum.

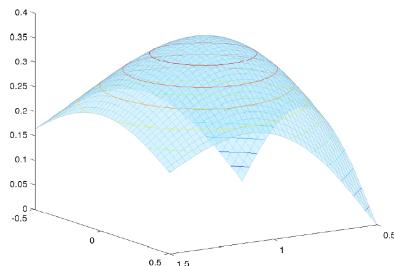


### Théorème 3.13 (Condition suffisante d'extrémum local dans un ouvert (cas de 2 variables))

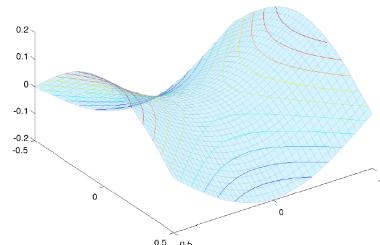
Soit  $f$  une fonction de classe  $\mathcal{C}^2$  sur un ouvert  $\mathcal{D} \subset \mathbb{R}^2$  et  $(x_0, y_0)$  un point stationnaire ; posons

$$\det(H_f(x_0, y_0)) = \partial_{xx} f(x_0, y_0) \cdot \partial_{yy} f(x_0, y_0) - (\partial_{xy} f(x_0, y_0))^2,$$

le déterminant de la matrice hessienne de  $f$  évalué en  $(x_0, y_0)$ .



(a) Point de maximum



(b) Point de selle



- ★ Si  $\det(H_f(x_0, y_0)) > 0$ , alors  $f$  présente un extrémum relatif en  $(x_0, y_0)$ ; il s'agit
  - \* d'un maximum si  $\partial_{xx}f(x_0, y_0) < 0$
  - \* d'un minimum si  $\partial_{xx}f(x_0, y_0) > 0$ ;
- \* si  $\det(H_f(x_0, y_0)) < 0$ , alors  $f$  présente un point-selle (ou point-col) en  $(x_0, y_0)$ ; ce n'est pas un extrémum;
- \* si  $\det(H_f(x_0, y_0)) = 0$ , on ne peut pas conclure à partir des dérivées secondes.

En résumé, si  $\partial_x f(x_0, y_0) = 0$  et  $\partial_y f(x_0, y_0) = 0$ , la nature du point critique  $(x_0, y_0)$  est déterminée par le tableau suivant :

$\det(H_f(x_0, y_0))$	$\partial_{xx}f(x_0, y_0)$	Nature de $(x_0, y_0)$
+	+	minimum local
+	-	maximum local
-		point-selle
0		on ne peut pas conclure

#### EXEMPLE

On veut étudier la fonction  $f(x, y) = x^2 + y^2 - 2x - 4y$  sur  $\mathbb{R}^2$ . Elle a pour dérivées partielles  $\partial_x f(x, y) = 2x - 2$  et  $\partial_y f(x, y) = 2y - 4$  qui ne s'annulent qu'en  $(1, 2)$ , seul point où il peut donc y avoir un extrémum local. On étudie directement le signe de la différence

$$d(h, k) = f(1 + h, 2 + k) - f(1, 2) = h^2 + k^2 > 0.$$

Comme cette différence est positive pour  $h$  et  $k$  voisins de 0 il s'agit d'un minimum. En effet,  $\partial_{xx}f(1, 2) = 2 > 0$ ,  $\partial_{yy}f(1, 2) = 2$ ,  $\partial_{xy}f(1, 2) = 0$  donc  $\det(H_f(1, 2)) = 4 > 0$  et il s'agit bien d'un minimum.

#### EXEMPLE

Pour déterminer les extrema libres de la fonction  $f(x, y) = x^2 + y^3 - 2xy - y$  dans  $\mathbb{R}^2$ , on constate d'abord que  $f$  est un polynôme, donc différentiable dans l'ouvert  $\mathbb{R}^2$ . Les seuls candidats extrema locaux sont les points critiques. Toutefois, nous ne disposons d'aucune garantie à priori sur le fait que les éventuels extrema locaux soient globaux.

**Recherche des points critiques** On a

$$\nabla f = \mathbf{0} \iff \begin{pmatrix} 2x - 2y \\ 3y^2 - 2x - 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff (x, y) = \left(-\frac{1}{3}, -\frac{1}{3}\right) \text{ ou } (x, y) = (1, 1).$$

Les deux candidats sont donc  $(-\frac{1}{3}, -\frac{1}{3})$  et  $(1, 1)$ .

**Classification** La matrice hessienne de  $f$  en un point  $(x, y) \in \mathbb{R}^2$  est

$$H_f(x, y) = \begin{pmatrix} \partial_{xx}f(x, y) & \partial_{xy}f(x, y) \\ \partial_{yx}f(x, y) & \partial_{yy}f(x, y) \end{pmatrix} = \begin{pmatrix} 2 & -2 \\ -2 & 6y \end{pmatrix}.$$

Comme  $\det(H_f(-\frac{1}{3}, -\frac{1}{3})) < 0$  et  $D(1, 1) > 0$ , alors  $(-\frac{1}{3}, -\frac{1}{3})$  est un point-selle et  $f$  admet en  $(1, 1)$  un minimum local de valeur  $f(1, 1) = -1$ . Ce minimum n'est cependant pas global puisque, par exemple,  $f(0, -2) = -6 < f(1, 1) = -1$ .

## 3.2 Application : fonction de meilleur approximation (*fitting*)

Lorsqu'un chercheur met au point une expérience (parce qu'il a quelques raisons de croire que les deux grandeurs  $x$  et  $y$  sont liées par une fonction  $f$ ), il récolte des données sous la forme de points  $\{(x_i, y_i)\}_{i=0}^n$  mais en général ces données sont affectées par des erreurs de mesure. Lorsqu'il en fait une représentation graphique il cherche  $f$  pour qu'elle s'ajuste le mieux possible aux points observés. Soit  $d_i = y_i - f(x_i)$  l'écart vertical du point  $(x_i, y_i)$  par rapport à la fonction  $f$ . La méthode des moindres carrés est celle qui choisit  $f$  de sorte que *la somme des carrés de ces déviations soit minimale*:

$$\text{minimiser} \quad \mathcal{E}_f = \sum_{i=0}^n (y_i - f(x_i))^2.$$

Le choix de la forme de  $f$  dépend du chercheur, on peut par exemple choisir :

- \*  $f$  affine, i.e.  $f(x) = a_0 + a_1 x$ , ainsi l'erreur est une fonction de deux variables et l'on a

$$\begin{aligned} \mathcal{E}(a_0, a_1) &= \sum_{i=0}^n (y_i - a_0 - a_1 x_i)^2 \\ \nabla \mathcal{E}(a_0, a_1) &= \begin{pmatrix} \partial_{a_0} \mathcal{E} \\ \partial_{a_1} \mathcal{E} \end{pmatrix} = \begin{pmatrix} -2 \sum_{i=0}^n x_i^0 (y_i - a_0 - a_1 x_i) \\ -2 \sum_{i=0}^n x_i^1 (y_i - a_0 - a_1 x_i) \end{pmatrix} \end{aligned}$$

- \*  $f$  polynomiale de degré  $m$ , i.e.  $f(x) = a_0 + a_1 x + \dots + a_m x^m = \sum_{j=0}^m a_j x^j$ , ainsi l'erreur est une fonction de  $m+1$  variables et l'on a

$$\begin{aligned} \mathcal{E}(a_0, a_1, a_2, \dots, a_m) &= \sum_{i=0}^n (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \dots - a_m x_i^m)^2 = \sum_{i=0}^n \left( y_i - \sum_{j=0}^m a_j x_i^j \right)^2 \\ \nabla \mathcal{E}(a_0, a_1, \dots, a_m) &= \begin{pmatrix} \partial_{a_0} \mathcal{E} \\ \partial_{a_1} \mathcal{E} \\ \partial_{a_2} \mathcal{E} \\ \vdots \\ \partial_{a_m} \mathcal{E} \end{pmatrix} = \begin{pmatrix} -2 \sum_{i=0}^n x_i^0 (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \dots - a_m x_i^m) \\ -2 \sum_{i=0}^n x_i^1 (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \dots - a_m x_i^m) \\ -2 \sum_{i=0}^n x_i^2 (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \dots - a_m x_i^m) \\ \vdots \\ -2 \sum_{i=0}^n x_i^m (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \dots - a_m x_i^m) \end{pmatrix} \end{aligned}$$

- \*  $f$  combinaison linéaire de  $m$  fonctions qui constituent une base d'un espace vectoriel, i.e.  $f(x) = a_0 \phi_0(x) + a_1 \phi_1(x) + \dots + a_m \phi_m(x) = \sum_{j=0}^m a_j \phi_j(x)$ , ainsi l'erreur est une fonction de  $m+1$  variables et l'on a

$$\begin{aligned} \mathcal{E}(a_0, a_1, a_2, \dots, a_m) &= \sum_{i=0}^n (y_i - a_0 \phi_0(x_i) - a_1 \phi_1(x_i) - a_2 \phi_2(x_i) - \dots - a_m \phi_m(x_i))^2 = \sum_{i=0}^n \left( y_i - \sum_{j=0}^m a_j \phi_j(x_i) \right)^2 \\ \nabla \mathcal{E}(a_0, a_1, \dots, a_m) &= \begin{pmatrix} \partial_{a_0} \mathcal{E} \\ \partial_{a_1} \mathcal{E} \\ \partial_{a_2} \mathcal{E} \\ \vdots \\ \partial_{a_m} \mathcal{E} \end{pmatrix} = \begin{pmatrix} -2 \sum_{i=0}^n \phi_0(x_i) (y_i - a_0 \phi_0(x_i) - a_1 \phi_1(x_i) - a_2 \phi_2(x_i) - \dots - a_m \phi_m(x_i)) \\ -2 \sum_{i=0}^n \phi_1(x_i) (y_i - a_0 \phi_0(x_i) - a_1 \phi_1(x_i) - a_2 \phi_2(x_i) - \dots - a_m \phi_m(x_i)) \\ -2 \sum_{i=0}^n \phi_2(x_i) (y_i - a_0 \phi_0(x_i) - a_1 \phi_1(x_i) - a_2 \phi_2(x_i) - \dots - a_m \phi_m(x_i)) \\ \vdots \\ -2 \sum_{i=0}^n \phi_m(x_i) (y_i - a_0 \phi_0(x_i) - a_1 \phi_1(x_i) - a_2 \phi_2(x_i) - \dots - a_m \phi_m(x_i)) \end{pmatrix} \end{aligned}$$

Bien évidemment, si  $\phi_j(x) = x^j$  on retrouve le cas de  $f$  polynomiale de degré  $m$ , mais ce n'est pas le seul choix possible. On peut par exemple choisir  $\phi_j(x) = e^{jx}$ ,  $\phi_j(x) = \cos(jx)$ ,  $\phi_j(x) = x^{-j}$ ...

### 3.2.1 Fitting par une relation affine

Supposons que les deux grandeurs  $x$  et  $y$  sont liées approximativement par une relation affine, i.e.  $f(x) = \sum_{j=0}^m \alpha_j x^j$  avec  $m=1$  et donc  $f(x) = \alpha_0 + \alpha_1 x$  pour certaines valeurs de  $\alpha_0$  et  $\alpha_1$  (autrement dit, lorsqu'on affiche ces points dans un plan cartésien, les points ne sont pas exactement alignés mais cela semble être dû à des erreurs de mesure). On souhaite alors trouver les constantes  $\alpha_0$  et  $\alpha_1$  pour que la droite d'équation  $y = \alpha_0 + \alpha_1 x$  s'ajuste *le mieux possible* aux points observés. Pour cela, introduisons  $d_i \equiv y_i - (\alpha_0 + \alpha_1 x_i)$  l'écart vertical du point  $(x_i, y_i)$  par rapport à la droite.

La méthode des moindres carrés est celle qui choisit  $\alpha_0$  et  $\alpha_1$  de sorte que *la somme des carrés de ces déviations soit minimale*. Pour cela, on doit minimiser la fonction  $\mathcal{E}: \mathbb{R}^2 \rightarrow \mathbb{R}_+$  définie par

$$\mathcal{E}(\alpha_0, \alpha_1) = \sum_{i=0}^n d_i^2 = \sum_{i=0}^n (y_i - \alpha_0 - \alpha_1 x_i)^2.$$

Pour minimiser  $\mathcal{E}$  on cherche d'abord les points stationnaires, i.e. les points  $(\alpha_0, \alpha_1)$  qui vérifient  $\frac{\partial \mathcal{E}}{\partial \alpha_0} = \frac{\partial \mathcal{E}}{\partial \alpha_1} = 0$ . Puisque

$$\frac{\partial \mathcal{E}}{\partial \alpha_0}(\alpha_0, \alpha_1) = -2 \left( \sum_{i=0}^n (y_i - \alpha_0 - \alpha_1 x_i) \right), \quad \frac{\partial \mathcal{E}}{\partial \alpha_1}(\alpha_0, \alpha_1) = -2 \left( \sum_{i=0}^n x_i (y_i - \alpha_0 - \alpha_1 x_i) \right),$$

alors

$$\begin{aligned} \begin{cases} \frac{\partial \mathcal{E}}{\partial \alpha_0}(\alpha_0, \alpha_1) = 0 \\ \frac{\partial \mathcal{E}}{\partial \alpha_1}(\alpha_0, \alpha_1) = 0 \end{cases} &\iff \begin{cases} \sum_{i=0}^n (y_i - \alpha_0 - \alpha_1 x_i) = 0 \\ \sum_{i=0}^n x_i (y_i - \alpha_0 - \alpha_1 x_i) = 0 \end{cases} \iff \begin{cases} \sum_{i=0}^n y_i - \alpha_0 \sum_{i=0}^n 1 - \alpha_1 \sum_{i=0}^n x_i = 0 \\ \sum_{i=0}^n x_i y_i - \alpha_0 \sum_{i=0}^n x_i - \alpha_1 \sum_{i=0}^n x_i^2 = 0 \end{cases} \\ &\iff \begin{cases} (n+1)\alpha_0 + (\sum_{i=0}^n x_i) \alpha_1 = \sum_{i=0}^n y_i \\ (\sum_{i=0}^n x_i) \alpha_0 + (\sum_{i=0}^n x_i^2) \alpha_1 = \sum_{i=0}^n y_i x_i \end{cases} \iff \underbrace{\begin{bmatrix} (n+1) & \sum_{i=0}^n x_i \\ \sum_{i=0}^n x_i & \sum_{i=0}^n x_i^2 \end{bmatrix}}_{\mathbb{F}} \underbrace{\begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix}}_{\mathbf{a}} = \underbrace{\begin{bmatrix} \sum_{i=0}^n y_i \\ \sum_{i=0}^n x_i y_i \end{bmatrix}}_{\mathbf{b}} \end{aligned}$$

Notons que l'élément  $(\mathbb{F})_{kj} = \sum_{i=0}^n x_i^{k+j}$  est le produit scalaire du vecteur  $(x_0^k, x_1^k)$  avec le vecteur  $(x_0^j, x_1^j)$  et que l'élément  $b_k = \sum_{i=0}^n x_i^k y_i$  est le produit scalaire du vecteur  $(x_0^k, x_1^k)$  avec le vecteur colonne  $\mathbf{y} = (y_0, y_1)$ ; on peut alors écrire  $\mathbb{F} = \mathbb{A}^T \mathbb{A}$  et  $\mathbf{b} = \mathbb{A}^T \mathbf{y}$  avec  $(\mathbb{A})_{ik} = x_i^k$  avec  $i = 0, \dots, n$  et  $k = 0, 1$ :

$$\mathbb{A} \stackrel{\text{def}}{=} \underbrace{\begin{pmatrix} 1 & x_0 \\ 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}}_{(n+1) \times (m+1)}.$$

En effet,

$$\mathbb{A}^T \mathbb{A} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_0 & x_1 & \dots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_0 \\ 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} n+1 & \sum_{i=0}^n x_i \\ \sum_{i=0}^n x_i & \sum_{i=0}^n x_i^2 \end{pmatrix} \quad \mathbb{A}^T \mathbf{y} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_0 & x_1 & \dots & x_n \end{pmatrix} \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^n y_i \\ \sum_{i=0}^n x_i y_i \end{pmatrix}$$

On peut résoudre à la main ce système linéaire et on trouve

$$\begin{cases} \alpha_0 = \frac{(\sum_{i=0}^n x_i)(\sum_{i=0}^n x_i y_i) - (\sum_{i=0}^n y_i)(\sum_{i=0}^n x_i^2)}{(\sum_{i=0}^n x_i)^2 - (n+1)(\sum_{i=0}^n x_i^2)}, \\ \alpha_1 = \frac{(\sum_{i=0}^n x_i)(\sum_{i=0}^n y_i) - (n+1)(\sum_{i=0}^n x_i y_i)}{(\sum_{i=0}^n x_i)^2 - (n+1)(\sum_{i=0}^n x_i^2)}. \end{cases}$$

On a trouvé un seul point stationnaire. La fonction étant convexe pour tout  $(\alpha_0, \alpha_1)$ , on peut conclure qu'il s'agit d'un minimum. Sinon, on peut vérifier la nature du point critique en étudiant la matrice Hessienne :

$$H_{\mathcal{E}}(\alpha_0, \alpha_1) = 2 \begin{pmatrix} (n+1) & \sum_{i=0}^n x_i \\ \sum_{i=0}^n x_i & \sum_{i=0}^n x_i^2 \end{pmatrix}$$

et  $\det(H_{\mathcal{E}}(\alpha_0, \alpha_1)) = 4 \left( (n+1) \sum_{i=1}^n x_i^2 - (\sum_{i=0}^n x_i)^2 \right) > 0$  avec  $\partial_{\alpha_0 \alpha_0} \mathcal{E}(\alpha_0, \alpha_1) > 0$  donc il s'agit bien d'un minimum.

### ✿ Remarque

Si on note

$$\bar{x} = \frac{1}{n+1} \sum_{i=0}^n x_i, \quad \bar{x^2} = \frac{1}{n+1} \sum_{i=0}^n x_i^2, \quad \bar{xy} = \frac{1}{n+1} \sum_{i=0}^n x_i y_i, \quad \bar{y} = \frac{1}{n+1} \sum_{i=0}^n y_i.$$

alors

$$\alpha_0 = \frac{\bar{y} \bar{x^2} - \bar{x} \bar{xy}}{\bar{x^2} - (\bar{x})^2}, \quad \alpha_1 = \frac{\bar{x} \bar{y} - \bar{xy}}{(\bar{x})^2 - \bar{x^2}}.$$

La droite d'équation  $y = \alpha_1 x + \alpha_0$  ainsi calculée s'appelle *droite de régression de y par rapport à x* et passe par le point moyen  $(\bar{x}, \bar{y})$ .

### 3.2.2 Fitting polynomial

On considère un ensemble de points expérimentaux  $\{(x_i, y_i)\}_{i=0}^n$  et on suppose que les deux grandeurs  $x$  et  $y$  sont liées, au moins approximativement, par une relation polynomiale, c'est-à-dire de la forme  $y = \sum_{j=0}^m a_j x^j$  pour certaines valeurs de  $a_j$ . On souhaite alors trouver les  $m+1$  constantes  $a_j$  pour que le polynôme d'équation  $f(x) = \sum_{j=0}^m a_j x^j$  s'ajuste le mieux possible aux points observés. Soit  $d_i(a_0, a_1, \dots, a_m) = y_i - (\sum_{j=0}^m a_j x_i^j)$  l'écart vertical du point  $(x_i, y_i)$  par rapport au polynôme. La méthode des moindres carrés est celle qui choisit les  $a_j$  de sorte que la somme des carrés de ces déviations soit minimale.

Pour cela, on doit minimiser la fonction  $\mathcal{E}: \mathbb{R}^{m+1} \rightarrow \mathbb{R}_+$  définie par

$$\mathcal{E}(a_0, a_1, a_2, \dots, a_m) = \sum_{i=0}^n (y_i - f(x_i))^2 = \sum_{i=0}^n \left( y_i - \sum_{j=0}^m a_j x_i^j \right)^2 = \sum_{i=0}^n (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \dots - a_m x_i^m)^2.$$

Pour minimiser  $\mathcal{E}$  on cherche d'abord ses points stationnaires, i.e. les points qui vérifient  $\frac{\partial \mathcal{E}}{\partial a_j} = 0$  pour  $j = 0, \dots, m$ . Puisque

$$\frac{\partial \mathcal{E}}{\partial a_0}(a_0, a_1, a_2, \dots, a_m) = -2 \sum_{i=0}^n x_i^0 \left( y_i - \sum_{j=0}^m a_j x_i^j \right) = -2 \sum_{i=0}^n x_i^0 (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \dots - a_m x_i^m),$$

$$\frac{\partial \mathcal{E}}{\partial a_1}(a_0, a_1, a_2, \dots, a_m) = -2 \sum_{i=0}^n x_i^1 \left( y_i - \sum_{j=0}^m a_j x_i^j \right) = -2 \sum_{i=0}^n x_i^1 (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \dots - a_m x_i^m),$$

$$\frac{\partial \mathcal{E}}{\partial a_2}(a_0, a_1, a_2, \dots, a_m) = -2 \sum_{i=0}^n x_i^2 \left( y_i - \sum_{j=0}^m a_j x_i^j \right) = -2 \sum_{i=0}^n x_i^2 (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \dots - a_m x_i^m),$$

⋮

$$\frac{\partial \mathcal{E}}{\partial a_m}(a_0, a_1, a_2, \dots, a_m) = -2 \sum_{i=0}^n x_i^m \left( y_i - \sum_{j=0}^m a_j x_i^j \right) = -2 \sum_{i=0}^n x_i^m (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \dots - a_m x_i^m),$$

on obtient alors le système linéaire  $\mathbb{F}\mathbf{a} = \mathbf{b}$  de  $(m+1)$  équations en les  $(m+1)$  inconnues  $a_0, a_1, a_2, \dots, a_m$  suivant

$$\begin{cases} \frac{\partial \mathcal{E}}{\partial a_0}(a_0, a_1, a_2, \dots, a_m) = 0 \\ \frac{\partial \mathcal{E}}{\partial a_1}(a_0, a_1, a_2, \dots, a_m) = 0 \\ \frac{\partial \mathcal{E}}{\partial a_2}(a_0, a_1, a_2, \dots, a_m) = 0 \\ \vdots \\ \frac{\partial \mathcal{E}}{\partial a_m}(a_0, a_1, a_2, \dots, a_m) = 0 \end{cases} \iff \begin{cases} a_0(n+1) + a_1 \sum_{i=0}^n x_i + a_2 \sum_{i=0}^n x_i^2 + \dots + a_m \sum_{i=0}^n x_i^m = \sum_{i=0}^n y_i \\ a_0 \sum_{i=0}^n x_i + a_1 \sum_{i=0}^n x_i^2 + a_2 \sum_{i=0}^n x_i^3 + \dots + a_m \sum_{i=0}^n x_i^{m+1} = \sum_{i=0}^n y_i x_i \\ a_0 \sum_{i=0}^n x_i^2 + a_1 \sum_{i=0}^n x_i^3 + a_2 \sum_{i=0}^n x_i^4 + \dots + a_m \sum_{i=0}^n x_i^{m+2} = \sum_{i=0}^n y_i x_i^2 \\ \vdots \\ a_0 \sum_{i=0}^n x_i^m + a_1 \sum_{i=0}^n x_i^{m+1} + a_2 \sum_{i=0}^n x_i^{m+2} + \dots + a_m \sum_{i=0}^n x_i^{2m} = \sum_{i=0}^n y_i x_i^m \end{cases} \iff \underbrace{\begin{bmatrix} (n+1) & \sum_{i=0}^n x_i & \sum_{i=0}^n x_i^2 & \dots & \sum_{i=0}^n x_i^m \\ \sum_{i=0}^n x_i & \sum_{i=0}^n x_i^2 & \sum_{i=0}^n x_i^3 & \dots & \sum_{i=0}^n x_i^{m+1} \\ \sum_{i=0}^n x_i^2 & \sum_{i=0}^n x_i^3 & \sum_{i=0}^n x_i^4 & \dots & \sum_{i=0}^n x_i^{m+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=0}^n x_i^m & \sum_{i=0}^n x_i^{m+1} & \sum_{i=0}^n x_i^{m+2} & \dots & \sum_{i=0}^n x_i^{2m} \end{bmatrix}}_{\mathbb{F}} \underbrace{\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}}_{\mathbf{a}} = \underbrace{\begin{bmatrix} \sum_{i=0}^n y_i \\ \sum_{i=0}^n y_i x_i \\ \sum_{i=0}^n y_i x_i^2 \\ \vdots \\ \sum_{i=0}^n y_i x_i^m \end{bmatrix}}_{\mathbf{b}}$$

Quand  $m \geq n$ , le polynôme de meilleure approximation coïncide avec le polynôme d'interpolation de  $\mathbb{R}_n[x]$ .

Notons que l'élément  $(\mathbb{F})_{kj} = \sum_{i=0}^n x_i^{k+j}$  est le produit scalaire du vecteur  $(x_0^k, x_1^k, \dots, x_n^k)$  avec le vecteur  $(x_0^j, x_1^j, \dots, x_n^j)$  et que l'élément  $b_k = \sum_{i=0}^n x_i^k y_i$  est le produit scalaire du vecteur  $(x_0^k, x_1^k, \dots, x_n^k)$  avec le vecteur  $\mathbf{y} = (y_0, y_1, \dots, y_n)$ ; on peut alors écrire  $\mathbb{F} = \mathbb{A}^T \mathbb{A}$  et  $\mathbf{b} = \mathbb{A}^T \mathbf{y}$  avec  $(\mathbb{A})_{ik} = x_i^k$  avec  $i = 0, \dots, n$  et  $k = 0, \dots, m$ :

$$\mathbb{A} \stackrel{\text{def}}{=} \underbrace{\begin{bmatrix} 1 & x_0 & \dots & x_0^m \\ 1 & x_1 & \dots & x_1^m \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \dots & x_n^m \end{bmatrix}}_{(n+1) \times (m+1)}.$$

On reconnaît une sous-matrice de la matrice de VANDERMONDE.

### Remarque

Le système des moindres carrés ci-dessus est mal conditionné (*i.e.* il est de plus en plus sensible aux erreurs d'arrondis à mesure que  $m$  augmente). On se limite habituellement à des polynômes de degré peu élevé.

### 3.2.3 Fitting non polynomial

Une généralisation de l'approximation au sens des moindres carrés consiste à utiliser dans  $d_i$  des fonctions  $f(x_i)$  qui ne sont pas des polynômes mais des fonctions d'un espace vectoriel  $\mathcal{V}$  engendré par  $m+1$  fonctions indépendantes  $\{\phi_j, j=0, \dots, m\}$ . On peut considérer par exemple des fonctions trigonométriques  $\phi_j(x) = \cos(jx)$ , des fonctions exponentielles  $\phi_j(x) = e^{jx}$  etc. Le choix des fonctions  $\{\phi_j\}$  est en pratique dicté par la forme supposée de la loi décrivant les données.

On considère un ensemble de points expérimentaux  $\{(x_i, y_i)\}_{i=0}^n$  et on suppose que les deux grandeurs  $x$  et  $y$  sont liées, au moins approximativement, par une relation de la forme  $y = \sum_{j=0}^m a_j \phi_j(x)$  pour certaines valeurs de  $a_j$  (le fitting polynomiale correspond à  $\phi_j(x) = x^j$ ). On souhaite alors trouver les  $m+1$  constantes  $a_j$  pour que la fonction d'équation  $y = \sum_{j=0}^m a_j \phi_j(x)$  s'ajuste le mieux possible aux points observés. Soit  $d_i(a_0, a_1, \dots, a_m) = y_i - (\sum_{j=0}^m a_j \phi_j(x_i))$  l'écart vertical du point  $(x_i, y_i)$  par rapport à cette fonction. La méthode des moindres carrés est celle qui choisit les  $a_j$  de sorte que la somme des carrés de ces déviations soit minimale.

Pour cela, on doit minimiser la fonction  $\mathcal{E}: \mathbb{R}^{m+1} \rightarrow \mathbb{R}_+$  définie par

$$\mathcal{E}(a_0, a_1, \dots, a_m) = \sum_{i=0}^n \left( y_i - \sum_{j=0}^m a_j \phi_j(x_i) \right)^2.$$

Pour minimiser  $\mathcal{E}$  on cherche d'abord ses points stationnaires, *i.e.* les points qui vérifient  $\frac{\partial \mathcal{E}}{\partial a_j} = 0$  pour  $j = 0, \dots, m$ . Puisque

$$\begin{aligned} \frac{\partial \mathcal{E}}{\partial a_0}(a_0, a_1, \dots, a_m) &= -2 \sum_{i=0}^n \left( \phi_0(x_i) \left( y_i - \sum_{j=0}^m a_j \phi_j(x_i) \right) \right), \\ \frac{\partial \mathcal{E}}{\partial a_1}(a_0, a_1, \dots, a_m) &= -2 \sum_{i=0}^n \left( \phi_1(x_i) \left( y_i - \sum_{j=0}^m a_j \phi_j(x_i) \right) \right), \\ &\vdots \\ \frac{\partial \mathcal{E}}{\partial a_m}(a_0, a_1, \dots, a_m) &= -2 \sum_{i=0}^n \left( \phi_m(x_i) \left( y_i - \sum_{j=0}^m a_j \phi_j(x_i) \right) \right), \end{aligned}$$

on obtient alors le système linéaire de  $(m+1)$  équations en les  $(m+1)$  inconnues  $a_0, a_1, \dots, a_m$  suivant

$$\begin{aligned} &\left\{ \begin{array}{l} \frac{\partial \mathcal{E}}{\partial a_0}(a_0, a_1, \dots, a_m) = 0 \\ \frac{\partial \mathcal{E}}{\partial a_1}(a_0, a_1, \dots, a_m) = 0 \\ \vdots \\ \frac{\partial \mathcal{E}}{\partial a_m}(a_0, a_1, \dots, a_m) = 0 \end{array} \right. \\ &\iff \left\{ \begin{array}{l} a_0 \sum_{i=0}^n \phi_0(x_i) \phi_0(x_i) + a_1 \sum_{i=0}^n \phi_0(x_i) \phi_1(x_i) + \dots + a_m \sum_{i=0}^n \phi_0(x_i) \phi_m(x_i) = \sum_{i=0}^n \phi_0(x_i) y_i \\ a_0 \sum_{i=0}^n \phi_1(x_i) \phi_0(x_i) + a_1 \sum_{i=0}^n \phi_1(x_i) \phi_1(x_i) + \dots + a_m \sum_{i=0}^n \phi_1(x_i) \phi_m(x_i) = \sum_{i=0}^n \phi_1(x_i) y_i \\ \vdots \\ a_0 \sum_{i=0}^n \phi_m(x_i) \phi_0(x_i) + a_1 \sum_{i=0}^n \phi_m(x_i) \phi_1(x_i) + \dots + a_m \sum_{i=0}^n \phi_m(x_i) \phi_m(x_i) = \sum_{i=0}^n \phi_m(x_i) y_i \end{array} \right. \\ &\iff \begin{pmatrix} \sum_{i=0}^n \phi_0(x_i) \phi_0(x_i) & \sum_{i=0}^n \phi_0(x_i) \phi_1(x_i) & \dots & \sum_{i=0}^n \phi_0(x_i) \phi_m(x_i) \\ \sum_{i=0}^n \phi_1(x_i) \phi_0(x_i) & \sum_{i=0}^n \phi_1(x_i) \phi_1(x_i) & \dots & \sum_{i=0}^n \phi_1(x_i) \phi_m(x_i) \\ \vdots & \vdots & & \vdots \\ \sum_{i=0}^n \phi_m(x_i) \phi_0(x_i) & \sum_{i=0}^n \phi_m(x_i) \phi_1(x_i) & \dots & \sum_{i=0}^n \phi_m(x_i) \phi_m(x_i) \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^n \phi_0(x_i) y_i \\ \sum_{i=0}^n \phi_1(x_i) y_i \\ \vdots \\ \sum_{i=0}^n \phi_m(x_i) y_i \end{pmatrix} \end{aligned}$$

Si on note  $\Phi_{kj} \stackrel{\text{def}}{=} \sum_{i=0}^n \phi_k(x_i) \phi_j(x_i)$ , on obtient alors le système linéaire  $\mathbb{F}\mathbf{a} = \mathbf{b}$  de  $(m+1)$  équations en les  $(m+1)$  inconnues

$a_0, a_1, \dots, a_m$  suivant

$$\underbrace{\begin{pmatrix} \Phi_{00} & \Phi_{01} & \dots & \Phi_{0m} \\ \Phi_{01} & \Phi_{11} & \dots & \Phi_{1m} \\ \vdots & \vdots & & \vdots \\ \Phi_{0m} & \Phi_{1m} & \dots & \Phi_{mm} \end{pmatrix}}_{\mathbb{F}} \underbrace{\begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{pmatrix}}_{\mathbf{a}} = \underbrace{\begin{pmatrix} \sum_{i=0}^n \phi_0(x_i) y_i \\ \sum_{i=0}^n \phi_1(x_i) y_i \\ \vdots \\ \sum_{i=0}^n \phi_m(x_i) y_i \end{pmatrix}}_{\mathbf{b}}$$

On remarque que si  $\phi_j(x) = x^j$  alors  $\Phi_{kj} = \sum_{i=0}^n x_i^{k+j}$  et on retrouve le cas du fitting polynomial.

Notons que l'élément  $\Phi_{kj}$  est le produit scalaire du vecteur  $(\phi_k(x_0), \phi_k(x_1), \dots, \phi_k(x_n))$  avec le vecteur  $(\phi_j(x_0), \phi_j(x_1), \dots, \phi_j(x_n))$  et que l'élément  $b_k = \sum_{i=0}^n \phi_k(x_i) y_i$  est le produit scalaire du vecteur  $(\phi_k(x_0), \phi_k(x_1), \dots, \phi_k(x_n))$  avec le vecteur colonne  $\mathbf{y} = (y_0, y_1, \dots, y_n)$ ; on peut alors écrire  $\mathbb{F} = \mathbb{A}^T \mathbb{A}$  et  $\mathbf{b} = \mathbb{A}^T \mathbf{y}$  avec  $(\mathbb{A})_{ik} = \phi_k(x_i)$  la matrice rectangulaire :

$$\mathbb{A} \stackrel{\text{def}}{=} \underbrace{\begin{pmatrix} \phi_0(x_0) & \phi_1(x_0) & \dots & \phi_m(x_0) \\ \phi_0(x_1) & \phi_1(x_1) & \dots & \phi_m(x_1) \\ \vdots & \vdots & & \vdots \\ \phi_0(x_n) & \phi_1(x_n) & \dots & \phi_m(x_n) \end{pmatrix}}_{(n+1) \times (m+1)}.$$

Le système linéaire carré  $\mathbb{A}^T \mathbb{A} \mathbf{a} = \mathbb{A}^T \mathbf{b}$  est équivalent au système linéaire rectangulaire  $\mathbb{A} \mathbf{a} = \mathbf{b}$ . Ce système peut être efficacement résolu avec la factorisation QR ou bien une décomposition en valeurs singulières de la matrice  $\mathbb{A}$ . Si  $n = m$  on trouve un système carré qui équivaut à la méthode directe d'interpolation.

**Fitting par un exponentiel** Soit  $a > 0$  et considérons la fonction  $f(x) = ae^{kx}$  : elle est non-linéaire mais si on prend son logarithme on obtient  $\ln(f(x)) = kx + \ln(a)$  qui est linéaire en  $k$  et a la forme  $mx + q$  avec  $m = k$  et  $q = \ln(a)$ . On peut alors faire une régression linéaire sur l'ensemble  $\{(x_i, \ln(y_i))\}_{i=0}^n$  et obtenir ainsi  $k$  et  $\ln(a)$ . Cependant ceci n'est pas équivalent à faire un fitting sur l'ensemble initial  $\{(x_i, y_i)\}_{i=0}^n$ . En effet, si on note  $d_i = y_i - ae^{kx_i}$  et  $D_i = \ln(y_i) - (kx_i + \ln(a))$ , lorsqu'on fait une régression linéaire sur l'ensemble  $\{(x_i, \ln(y_i))\}_{i=0}^n$  on minimise  $D_i$  et non  $d_i$ .

### EXEMPLE

Considérons l'ensemble de 3 points  $\{(-2, 4), (0, 0), (1, 1)\}$  (donc  $n = 2$ ). On se propose de calculer les fonctions de meilleures approximations avec

1.  $f(x) = a_0 + a_1 x$  ( $m = 1$  et  $\phi_j(x) = x^j$  avec  $j = 0, 1$ )
2.  $f = a_0 + a_1 x + a_2 x^2$  ( $m = 2$  et  $\phi_j(x) = x^j$  avec  $j = 0, 1, 2$ )
3.  $f(x) = a_0 + a_1 e^x$  ( $m = 1$  et  $\phi_j(x) = e^{jx}$  avec  $j = 0, 1$ )
4.  $f(x) = a_0 + a_1 e^x + a_2 e^{2x}$  ( $m = 2$  et  $\phi_j(x) = e^{jx}$  avec  $j = 0, 1, 2$ )

Posons les systèmes linéaires :

1. Pour  $m = 1$ , il s'agit de chercher  $a_0$  et  $a_1$  qui minimisent l'erreur  $\mathcal{E}(a_0, a_1) = \sum_{i=0}^2 (y_i - (a_0 + a_1 x_i))^2$ . Cela impose la résolution du système linéaire

$$\begin{pmatrix} (n+1) & \sum_{i=0}^n x_i \\ \sum_{i=0}^n x_i & \sum_{i=0}^n x_i^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^n y_i \\ \sum_{i=0}^n y_i x_i \end{pmatrix} \Rightarrow \begin{pmatrix} 3 & -1 \\ -1 & 5 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 5 \\ -7 \end{pmatrix}$$

Donc  $a_0 = \frac{27}{21}$  et  $a_1 = -\frac{8}{7}$ .

2. Pour  $m = 2$ , il s'agit de chercher  $a_0, a_1$  et  $a_2$  qui minimisent l'erreur  $\mathcal{E}(a_0, a_1, a_2) = \sum_{i=0}^2 (y_i - (a_0 + a_1 x_i + a_2 x_i^2))^2$ . Cela impose la résolution du système linéaire

$$\begin{pmatrix} (n+1) & \sum_{i=0}^n x_i & \sum_{i=0}^n x_i^2 \\ \sum_{i=0}^n x_i & \sum_{i=0}^n x_i^2 & \sum_{i=0}^n x_i^3 \\ \sum_{i=0}^n x_i^2 & \sum_{i=0}^n x_i^3 & \sum_{i=0}^n x_i^4 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^n y_i \\ \sum_{i=0}^n y_i x_i \\ \sum_{i=0}^n y_i x_i^2 \end{pmatrix} \text{ i.e. } \begin{pmatrix} 3 & -1 & 5 \\ -1 & 5 & -7 \\ 5 & -7 & 17 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 5 \\ -7 \\ 17 \end{pmatrix}$$

Donc  $a_0 = a_1 = 0$  et  $a_2 = 1$ , i.e.  $f(x) = x^2$ . Notons que dans ce cas  $\mathcal{E}(0, 0, 1) = 0$  : en effet  $m = n - 1$  et le fitting retrouve le polynôme d'interpolation.

3. Pour  $m = 1$ , il s'agit de chercher  $a_0$  et  $a_1$  qui minimisent l'erreur  $\mathcal{E}(a_0, a_1) = \sum_{i=0}^2 (y_i - a_0 - a_1 e^{x_i})^2$ . Cela impose la résolution du système linéaire

$$\begin{pmatrix} \sum_{i=0}^n \phi_0(x_i) \phi_0(x_i) & \sum_{i=0}^n \phi_0(x_i) \phi_1(x_i) \\ \sum_{i=0}^n \phi_0(x_i) \phi_1(x_i) & \sum_{i=0}^n \phi_1(x_i) \phi_1(x_i) \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^n \phi_0(x_i) y_i \\ \sum_{i=0}^n \phi_1(x_i) y_i \end{pmatrix} \Rightarrow \begin{pmatrix} \sum_{i=0}^n 1 & \sum_{i=0}^n e^{x_i} \\ \sum_{i=0}^n e^{x_i} & \sum_{i=0}^n e^{2x_i} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^n y_i \\ \sum_{i=0}^n e^{x_i} y_i \end{pmatrix}$$

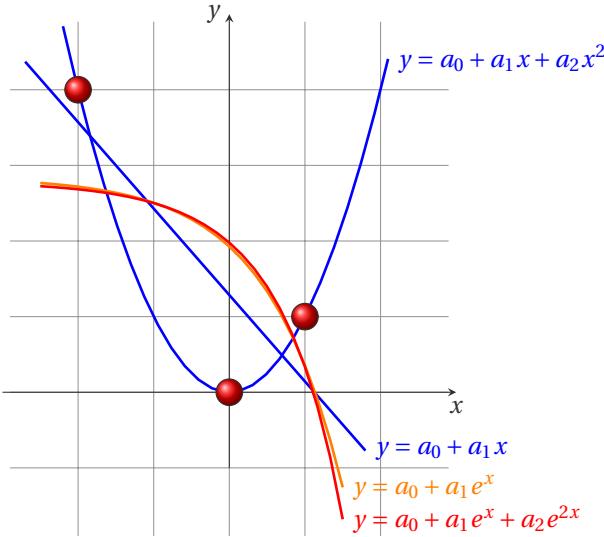
$$\Rightarrow \begin{pmatrix} 3 & e^{-2} + 1 + e \\ e^{-2} + 1 + e & e^{-4} + 1 + e^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 5 \\ 4e^{-2} + 0 + e \end{pmatrix}$$

Donc  $a_0 \approx 2.842$  et  $a_1 \approx -0.915$ .

4. Pour  $m = 2$ , il s'agit de chercher  $a_0$ ,  $a_1$  et  $a_2$  qui minimisent l'erreur  $\mathcal{E}(a_0, a_1, a_2) = \sum_{i=0}^2 (y_i - a_0 - a_1 e^{x_i} - a_2 e^{2x_i})^2$ . Cela impose la résolution du système linéaire

$$\begin{aligned} \begin{pmatrix} \sum_{i=0}^n \phi_0(x_i) \phi_0(x_i) & \sum_{i=0}^n \phi_0(x_i) \phi_1(x_i) & \sum_{i=0}^n \phi_0(x_i) \phi_2(x_i) \\ \sum_{i=0}^n \phi_1(x_i) \phi_0(x_i) & \sum_{i=0}^n \phi_1(x_i) \phi_1(x_i) & \sum_{i=0}^n \phi_1(x_i) \phi_2(x_i) \\ \sum_{i=0}^n \phi_2(x_i) \phi_0(x_i) & \sum_{i=0}^n \phi_2(x_i) \phi_1(x_i) & \sum_{i=0}^n \phi_2(x_i) \phi_2(x_i) \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} &= \begin{pmatrix} \sum_{i=0}^n \phi_0(x_i) y_i \\ \sum_{i=0}^n \phi_1(x_i) y_i \\ \sum_{i=0}^n \phi_2(x_i) y_i \end{pmatrix} \\ \Rightarrow \begin{pmatrix} \sum_{i=0}^n 1 & \sum_{i=0}^n e^{x_i} & \sum_{i=0}^n e^{2x_i} \\ \sum_{i=0}^n e^{x_i} & \sum_{i=0}^n e^{2x_i} & \sum_{i=0}^n e^{3x_i} \\ \sum_{i=0}^n e^{2x_i} & \sum_{i=0}^n e^{3x_i} & \sum_{i=0}^n e^{4x_i} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} &= \begin{pmatrix} \sum_{i=0}^n y_i \\ \sum_{i=0}^n e^{x_i} y_i \\ \sum_{i=0}^n e^{2x_i} y_i \end{pmatrix} \\ \Rightarrow \begin{pmatrix} 3 & e^{-2} + 1 + e & e^{-4} + 1 + e^2 \\ e^{-2} + 1 + e & e^{-4} + 1 + e^2 & e^{-6} + 1 + e^3 \\ e^{-4} + 1 + e^2 & e^{-6} + 1 + e^3 & e^{-8} + 1 + e^4 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} &= \begin{pmatrix} 5 \\ 4e^{-2} + 0 + e \\ 4e^{-4} + 0 + e^2 \end{pmatrix} \end{aligned}$$

Donc  $a_0 \approx 2.787$ ,  $a_1 \approx -0.755$  et  $a_2 \approx -0.054$ .



### EXEMPLE

Considérons l'ensemble de 3 points  $\{(1, 2), (2, 0), (3, -1)\}$  (donc  $n = 2$ ). On se propose de calculer les fonctions de meilleures approximations avec

1.  $f(x) = a_0 + a_1 \frac{1}{x}$  ( $m = 1$  et  $\phi_j(x) = x^{-j}$  avec  $j = 0, 1$ )
2.  $f(x) = a_0 + a_1 \frac{1}{x} + a_2 \frac{1}{x^2}$  ( $m = 2$  et  $\phi_j(x) = x^{-j}$  avec  $j = 0, 1, 2$ )

Posons les systèmes linéaires :

1. Pour  $m = 1$ , il s'agit de chercher  $a_0$  et  $a_1$  qui minimisent l'erreur  $\mathcal{E}(a_0, a_1) = \sum_{i=0}^2 (y_i - a_0 - a_1 \frac{1}{x_i})^2$ . Cela impose la résolution du système linéaire

$$\begin{pmatrix} \sum_{i=0}^n \phi_0(x_i) \phi_0(x_i) & \sum_{i=0}^n \phi_0(x_i) \phi_1(x_i) \\ \sum_{i=0}^n \phi_0(x_i) \phi_1(x_i) & \sum_{i=0}^n \phi_1(x_i) \phi_1(x_i) \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^n \phi_0(x_i) y_i \\ \sum_{i=0}^n \phi_1(x_i) y_i \end{pmatrix} \Rightarrow \begin{pmatrix} \sum_{i=0}^n 1 & \sum_{i=0}^n \frac{1}{x_i} \\ \sum_{i=0}^n \frac{1}{x_i} & \sum_{i=0}^n \frac{1}{x_i^2} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^n y_i \\ \sum_{i=0}^n \frac{1}{x_i} y_i \end{pmatrix} \Rightarrow \begin{pmatrix} 3 & \frac{11}{6} \\ \frac{11}{6} & \frac{49}{36} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{5}{3} \end{pmatrix}$$

Donc  $a_0 \approx -4.2308$  et  $a_1 \approx 6.9231$ .

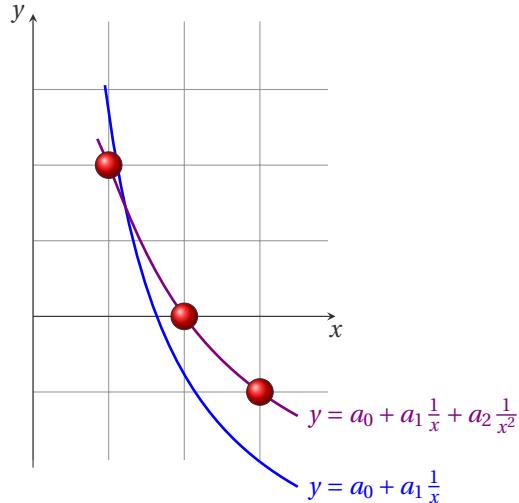
2. Pour  $m = 2$ , il s'agit de chercher  $a_0$ ,  $a_1$  et  $a_2$  qui minimisent l'erreur  $\mathcal{E}(a_0, a_1, a_2) = \sum_{i=0}^2 (y_i - a_0 - a_1 \frac{1}{x_i} - a_2 \frac{1}{x_i^2})^2$ . Cela impose la résolution du système linéaire

$$\begin{pmatrix} \sum_{i=0}^n \phi_0(x_i) \phi_0(x_i) & \sum_{i=0}^n \phi_0(x_i) \phi_1(x_i) & \sum_{i=0}^n \phi_0(x_i) \phi_2(x_i) \\ \sum_{i=0}^n \phi_1(x_i) \phi_0(x_i) & \sum_{i=0}^n \phi_1(x_i) \phi_1(x_i) & \sum_{i=0}^n \phi_1(x_i) \phi_2(x_i) \\ \sum_{i=0}^n \phi_2(x_i) \phi_0(x_i) & \sum_{i=0}^n \phi_2(x_i) \phi_1(x_i) & \sum_{i=0}^n \phi_2(x_i) \phi_2(x_i) \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^n \phi_0(x_i) y_i \\ \sum_{i=0}^n \phi_1(x_i) y_i \\ \sum_{i=0}^n \phi_2(x_i) y_i \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} \sum_{i=0}^n 1 & \sum_{i=0}^n \frac{1}{x_i} & \sum_{i=0}^n \frac{1}{x_i^2} \\ \sum_{i=0}^n \frac{1}{x_i} & \sum_{i=0}^n \frac{1}{x_i^2} & \sum_{i=0}^n \frac{1}{x_i^3} \\ \sum_{i=0}^n \frac{1}{x_i^2} & \sum_{i=0}^n \frac{1}{x_i^3} & \sum_{i=0}^n \frac{1}{x_i^4} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^n y_i \\ \sum_{i=0}^n \frac{1}{x_i} y_i \\ \sum_{i=0}^n \frac{1}{x_i^2} y_i \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} 3 & \frac{11}{6} & \frac{49}{36} \\ \frac{11}{6} & \frac{49}{36} & \frac{251}{216} \\ \frac{49}{36} & \frac{251}{216} & \frac{1393}{1296} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 1 \\ \frac{5}{3} \\ \frac{17}{9} \end{pmatrix}$$

Donc  $a_0 = -\frac{7}{2}$ ,  $a_1 = \frac{17}{2}$  et  $a_2 = -3$ .





## Exercices



### Optimisation de fonctions de plusieurs variables

#### Exercice 3.1

Déterminer les courbes de niveau des fonctions suivantes :

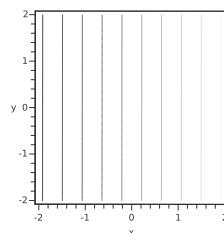
$$f(x, y) = x, \quad f(x, y) = y + 1, \quad f(x, y) = x + y - 1, \quad f(x, y) = e^{y-x^2}, \quad f(x, y) = y - \cos(x).$$

Esquissez ensuite leurs graphes (le graphe peut être vu comme un empilement de courbes de niveau qui forment une surface dans  $\mathbb{R}^3$ ).

#### Correction

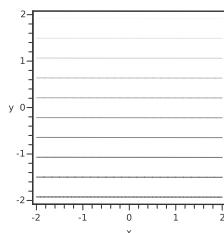
\*  $f(x, y) = x$  :

$f(x, y) = \kappa$ ssi  $x = \kappa$ , les courbes de niveau sont des droites verticales et la surface représentative de  $f$  est un plan.



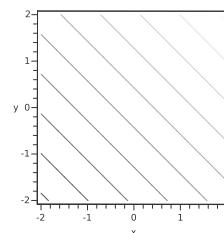
\*  $f(x, y) = y + 1$  :

$f(x, y) = \kappa$ ssi  $y = \kappa - 1$ , les courbes de niveau sont des droites horizontales et la surface représentative de  $f$  est un plan.



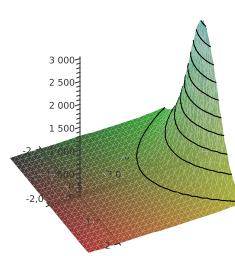
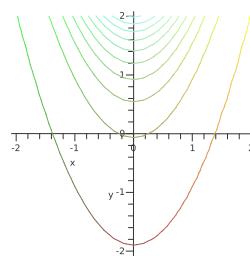
\*  $f(x, y) = x + y - 1$  :

$f(x, y) = \kappa$ ssi  $y = -x + (\kappa + 1)$ , les courbes de niveau sont des droites de pente  $-1$  et la surface représentative de  $f$  est un plan.



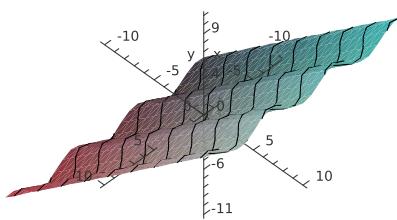
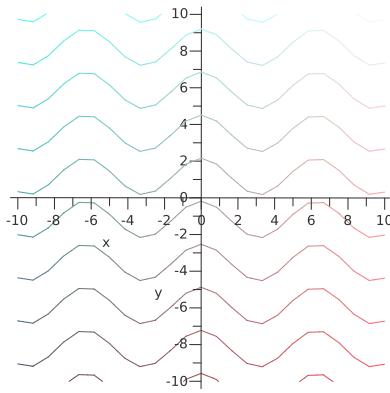
\*  $f(x, y) = e^{y-x^2}$  :

$f(x, y) = \kappa$ ssi  $y = x^2 + \ln(\kappa)$ , les courbes de niveau sont des paraboles. On observe notamment la croissance exponentielle marquée lorsque les valeurs prises par  $y$  sont grandes et celles prises par  $|x|$  sont petites.



\*  $f(x, y) = y - \cos(x)$  :

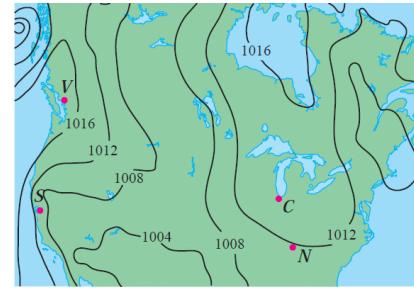
$f(x, y) = \kappa$ ssi  $y = \cos(x) + \kappa$



### Exercice 3.2

Dans la figure ci-contre on a tracé les isobares de l'Amérique du Nord au 12 août 2008. La pression indiquée est mesurée en millibars (mbar).

1. Donner une estimation de la pression
  - \* à Nashville (point N),
  - \* à Chicago (point C),
  - \* à San Francisco (point S)
  - \* et à Vancouver (point V).
2. Dans quelle ville le vent est le plus fort ?



### Correction

1. \* Au point N la pression est de 1012 mbar environ,  
   \* au point C la pression est de 1013 mbar environ,  
   \* au point S la pression est de 1010 mbar environ,  
   \* au point V la pression est comprise entre 1016 mbar et 1020 mbar ou entre 1012 mbar et 1016 mbar.
2. Le vent est plus fort à San Francisco car les lignes de pression sont le plus rapprochées.

### Exercice 3.3

Associer chaque fonction (1-6) à sa surface (A-F) et à ses courbes de niveau (I-VI) :

$$(1) f(x, y) = \sin(xy)$$

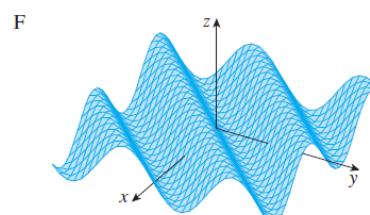
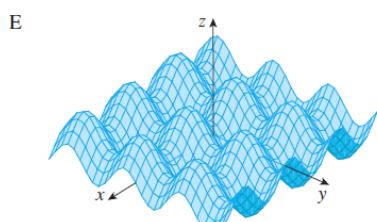
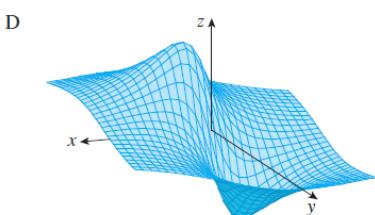
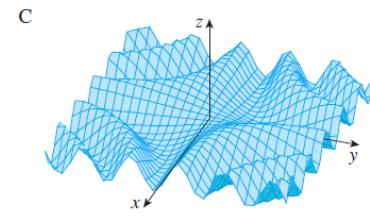
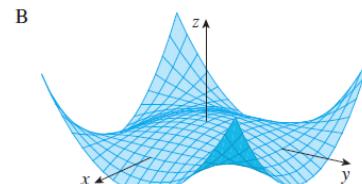
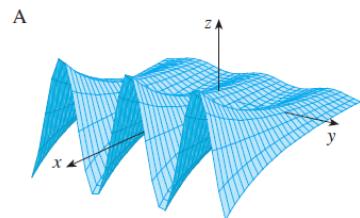
$$(2) f(x, y) = \sin(x - y)$$

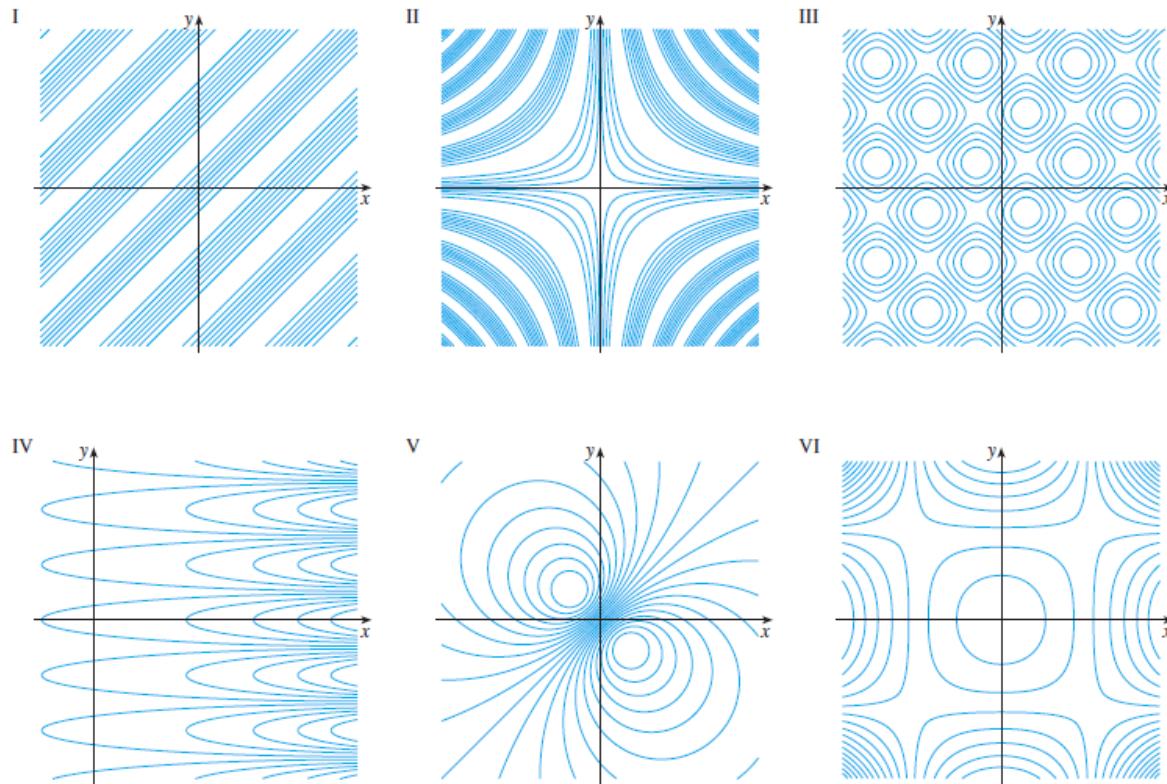
$$(3) f(x, y) = (1 - x^2)(1 - y^2)$$

$$(4) f(x, y) = \frac{x - y}{1 + x^2 + y^2}$$

$$(5) f(x, y) = e^x \cos(y)$$

$$(6) f(x, y) = \sin(x) - \sin(y)$$

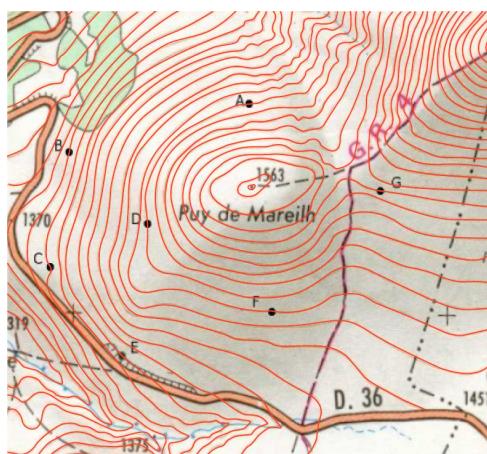




### Correction

1. C-II : la fonction est périodique en  $x$  et en  $y$ ;  $f$  ne change pas quand on échange  $x$  et  $y$ , i.e. le graphe est symétrique par rapport au plan d'équation  $y = x$ ;  $f(0, y) = f(x, 0) = 0$ .
2. F-I : la fonction est périodique en  $x$  et en  $y$ ;  $f$  est constante si  $y = x + \kappa$ .
3. B-VI :  $f(\pm 1, y) = f(x, \pm 1) = 0$ ; la trace dans le plan  $xz$  est  $z = 1 - x^2$  et dans le plan  $yz$  est  $z = 1 - y^2$ .
4. D-V :  $f(x, x) = 0$ ;  $f(x, y) > 0$  si  $x > y$ ;  $f(x, y) < 0$  si  $x < y$ .
5. A-IV : la fonction est périodique en  $y$ ;
6. E-III : la fonction est périodique en  $x$  et en  $y$ .

### Exercice 3.4 (Cartes topographiques du relief)



Sur une *carte topographique*, les courbes de niveau désignent les points de même altitude. On observe sur l'extrait de carte ci-contre de l'institut Géographique National (IGN), des courbes qui donnent une idée du relief (Massif du Sancy). Elles représentent des coupes horizontales successives du terrain à des altitudes qui varient de 10 mètres en 10 mètres. Tous les points de même altitude sont situés sur la même courbe de niveau.

1. Compléter le tableau

Point	A	B	C	D	E	F	G
Altitude		1370					

2. Lorsque les courbes de niveau se resserrent, que peut-on dire du relief?

3. La rivière coule-t-elle d'est en ouest ou vice-versa ?

### Correction

1. On a

Point	A	B	C	D	E	F	G
Altitude	1470	1370	1380	1470	1400	1460	1520

2. Les endroits du relief où les pentes sont plus escarpées ou plus douces correspondent respectivement aux courbes de niveau très rapprochées ou très distantes.  
 3. La rivière coule de l'est à l'ouest.

### Exercice 3.5

Calculer toutes les dérivées partielles d'ordre 1 des fonctions données :

1.  $f(x, y) = y^5 - 3xy$
2.  $f(x, y) = x^2 + 3xy^2 - 6y^5$
3.  $f(x, y) = x \cos(e^{xy})$
4.  $f(x, y) = x/y$
5.  $f(x, y) = x^y$
6.  $f(x, y, z) = x \cos(xz) + \ln(2 - \sin^2(y + z))$
7.  $f(x, t) = e^{-t} \cos(\pi x)$
8.  $z = (2x + 3y)^{10}$
9.  $f(x, y) = \frac{ax+by}{cx+dy}$
10.  $F(x, y) = \int_y^x \cos(e^t) dt$

### Correction

1.  $\partial_x f(x, y) = -3y$  et  $\partial_y f(x, y) = 5y^4 - 3x$
  2.  $\partial_x f(x, y) = 2x + 3y^2$  et  $\partial_y f(x, y) = 6xy - 30y^4$
  3.  $\partial_x f(x, y) = \cos(e^{xy}) - xye^{xy} \sin(e^{xy})$  et  $\partial_y f(x, y) = -x^2 e^{xy} \sin(e^{xy})$
  4.  $\partial_x f(x, y) = 1/y$  et  $\partial_y f(x, y) = -x/y^2$
  5.  $\partial_x f(x, y) = yx^y/x$  et  $\partial_y f(x, y) = \ln(x)x^y$
  6.  $\partial_x f(x, y, z) = \cos(xz) - xz \sin(xz)$ ,  $\partial_y f(x, y) = \frac{-2 \sin(y+z) \cos(y+z)}{2 - \sin^2(y+z)}$  et  $\partial_z f(x, y) = -x^2 \sin(xz) + \frac{-2 \sin(y+z) \cos(y+z)}{2 - \sin^2(y+z)}$
  7.  $\partial_x f(x, t) = -\pi e^{-t} \sin(\pi x)$  et  $\partial_t f(x, t) = -e^{-t} \cos(\pi x)$
  8.  $\partial_x z(x, y) = 20(2x + 3y)^9$  et  $\partial_y z(x, y) = 30(2x + 3y)^9$
  9.  $\partial_x f(x, y) = \frac{(ad-bc)y}{(cx+dy)^2}$  et  $\partial_y f(x, y) = \frac{(bc-ad)x}{(cx+dy)^2}$
  10.  $\partial_x F(x, y) = \cos(e^x)$  et  $\partial_y F(x, y) = -\cos(e^y)$
- $\triangle x^y = e^{y \ln(x)}$  donc  $x > 0$

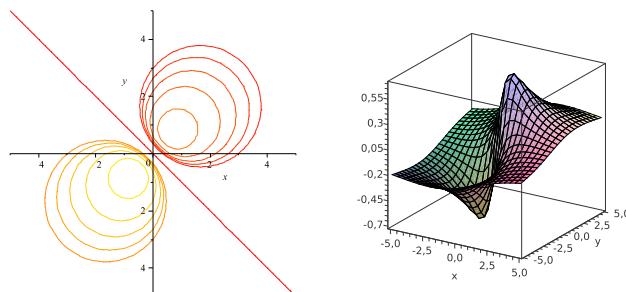
### Exercice 3.6

Soit  $f: \mathbb{R}^2 \rightarrow \mathbb{R}$  la fonction définie par  $f(x, y) = \frac{x+y}{1+x^2+y^2}$ .

1. Déterminer et représenter ses courbes de niveau.
2. Calculer ses dérivées partielles premières.
3. Écrire l'équation du plan tangent à  $f$  en  $(0, 0)$ .

### Correction

1. Les courbes de niveau de  $f$  sont les courbes d'équation  $f(x, y) = k$ , i.e. la droite d'équation  $y = -x$  pour  $k = 0$  et les courbes d'équation  $x^2 + y^2 - \frac{1}{k}x - \frac{1}{k}y + 1 = 0$  pour  $0 < k^2 < 1/2$  qui sont des cercles de centre  $(\frac{1}{2k}, \frac{1}{2k})$  et rayon  $\sqrt{\frac{1}{2k^2} - 1}$ .



2. Les deux dérivées premières partielles de  $f$  sont

$$\partial_x f(x, y) = \frac{1 - x^2 - 2xy + y^2}{(1 + x^2 + y^2)^2}, \quad \partial_y f(x, y) = \frac{1 + x^2 - 2xy - y^2}{(1 + x^2 + y^2)^2}.$$

3. L'équation du plan tangent à  $f$  en  $(0, 0)$  est

$$z = f(0, 0) + x\partial_x(0, 0) + y\partial_y(0, 0) = x + y.$$

### Exercice 3.7

Soit  $f: \mathbb{R}^2 \rightarrow \mathbb{R}$  une fonction de classe  $\mathcal{C}^2(\mathbb{R}^2)$  et  $(a, b)$  un point de  $\mathbb{R}^2$ . On suppose que

$$f(a, b) = 0, \quad \partial_x f(a, b) = 0, \quad \partial_y f(a, b) = 0, \quad \partial_{xx} f(a, b) = 1, \quad \partial_{yy} f(a, b) = 2, \quad \partial_{xy} f(a, b) = 3.$$

Le point  $(a, b)$  est-il un point critique ? Si oui, de quelle nature ?

#### Correction

Il est un point critique et plus particulièrement il s'agit d'un point-selle car  $\det(H_f(a, b)) < 0$ .

### Exercice 3.8

On suppose que  $(1, 1)$  est un point critique d'une fonction  $f$  dont les dérivées seconde sont continues. Dans chaque cas, que peut-on dire au sujet de  $f$  ?

1.  $\partial_{xx} f(1, 1) = 4, \partial_{xy} f(1, 1) = 1, \partial_{yy} f(1, 1) = 2$  ;

2.  $\partial_{xx} f(1, 1) = 4, \partial_{xy} f(1, 1) = 3, \partial_{yy} f(1, 1) = 2$ .

#### Correction

1. D'abord on calcule  $\det(H_f(1, 1)) = \partial_{xx} f(1, 1)\partial_{yy} f(1, 1) - (\partial_{xy} f(1, 1))^2 = 7$ . Comme  $\det(H_f(1, 1)) > 0$  et  $\partial_{xx} f(1, 1) > 0$ ,  $f$  a un minimum local en  $(1, 1)$ .

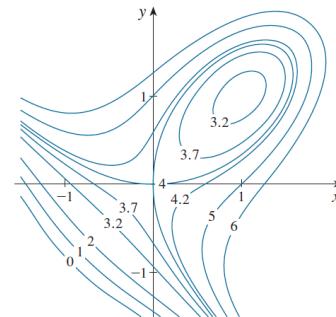
2. D'abord on calcule  $\det(H_f(1, 1)) = \partial_{xx} f(1, 1)\partial_{yy} f(1, 1) - (\partial_{xy} f(1, 1))^2 = -1$ . Comme  $\det(H_f(1, 1)) < 0$ ,  $f$  a un point-selle en  $(1, 1)$ .

### Exercice 3.9

À partir de la carte des courbes de niveau de la figure ci-contre, localiser les points critiques de  $f: \mathbb{R}^2 \rightarrow \mathbb{R}$  et préciser pour chacun de ces points s'il s'agit d'un point-selle ou d'un maximum ou d'un minimum local.

Vérifier ensuite le raisonnement sachant que

$$f(x, y) = 4 + x^3 + y^3 - 3xy.$$



#### Correction

Dans la figure, le point  $(1, 1)$  est entouré par des courbes de niveau qui sont de forme ovale et qui indiquent que si nous nous éloignons du point dans n'importe quelle direction les valeurs de  $f$  augmentent. Ainsi on pourrait s'attendre à un minimum local en ou à proximité de  $(1, 1)$ .

Les courbes de niveau proches du point  $(0, 0)$  ressemblent à des hyperboles, et si nous nous éloignons de l'origine, les valeurs de  $f$  augmentent dans certaines directions et diminuent dans d'autres, donc nous nous attendons à trouver un point selle. Vérifions cette analyse :

**Points critiques :**  $\partial_x f(x, y) = 3x^2 - 3y, \partial_y f(x, y) = 3y^2 - 3x$ . On a un point critique si les deux dérivées partielles s'annulent en même temps ; on trouve deux points critiques :  $(1, 1)$  et  $(0, 0)$ .

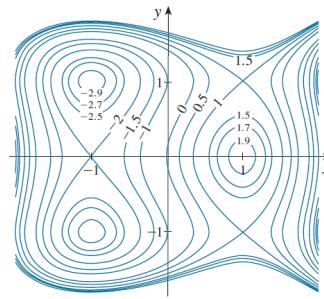
**Études des points critiques :** les dérivées secondes sont  $\partial_{xx} f(x, y) = 6x, \partial_{xy} f(x, y) = -3, \partial_{yy} f(x, y) = 6y$ , ainsi  $\det(H_f(x, y)) = \partial_{xx} f(x, y)\partial_{yy} f(x, y) - (\partial_{xy} f(x, y))^2 = 36xy - 9$ . Comme  $\det(H_f(1, 1)) > 0$  et  $\partial_{xx} f(1, 1) > 0$ ,  $f$  a un minimum local en  $(1, 1)$ . Comme  $\det(H_f(0, 0)) < 0$ ,  $f$  a un point-selle en  $(0, 0)$ .

### Exercice 3.10

À partir de la carte des courbes de niveau de la figure ci-contre, localiser les points critiques de  $f: \mathbb{R}^2 \rightarrow \mathbb{R}$  et préciser pour chacun de ces points s'il s'agit d'un point-selle ou d'un maximum ou d'un minimum local.

Vérifier ensuite le raisonnement sachant que

$$f(x, y) = 3x - x^3 - 2y^2 + y^4.$$



#### Correction

Dans la figure, les points  $(-1, -1)$  et  $(-1, 1)$  sont entourés par des courbes de niveau qui sont de forme ovale et qui indiquent que si nous nous éloignons du point dans n'importe quelle direction les valeurs de  $f$  augmentent. Ainsi on pourrait s'attendre à des minima locaux en ou à proximité de  $(-1, \pm 1)$ .

De la même manière, le point  $(1, 0)$  est entouré par des courbes de niveau qui sont de forme ovale et qui indiquent que si nous nous éloignons du point dans n'importe quelle direction les valeurs de  $f$  diminuent. Ainsi on pourrait s'attendre à un maximum local en ou à proximité de  $(1, 0)$ .

Les courbes de niveau proche des points  $(-1, 0)$ ,  $(1, 1)$  et  $(1, -1)$  ressemblent à des hyperboles, et si nous nous éloignons de ces points, les valeurs de  $f$  augmentent dans certaines directions et diminuent dans d'autres, donc nous nous attendons à trouver des points de selle.

Vérifions cette analyse :

$$\nabla f = \mathbf{0} \iff \begin{cases} 3 - 3x^2 = 0 \\ -4y + 4y^3 = 0 \end{cases}$$

donc les points critiques sont  $(1, 0)$ ,  $(1, 1)$ ,  $(1, -1)$ ,  $(-1, 0)$ ,  $(-1, 1)$ ,  $(-1, -1)$ . Les dérivées secondes sont  $\partial_{xx}f(x, y) = -6x$ ,  $\partial_{xy}f(x, y) = 0$ ,  $\partial_{yy}f(x, y) = 12y^2 - 4$ , ainsi  $\det(H_f(x, y)) = \partial_{xx}f(x, y)\partial_{yy}f(x, y) - (\partial_{xy}f(x, y))^2 = -72xy^2 + 24x$ .

Point critique $(x_0, y_0)$	$\det(H_f(x_0, y_0))$	$\partial_{xx}f(x_0, y_0)$	Conclusion
$(1, 0)$	$24 > 0$	$-6 < 0$	$f$ a un maximum local en $(1, 0)$
$(1, 1)$	$-48 < 0$		$f$ a un point-selle en $(1, 1)$
$(1, -1)$	$-48 < 0$	$-6 < 0$	$f$ a un point-selle en $(1, -1)$
$(-1, 0)$	$-24 < 0$		$f$ a un point-selle en $(-1, 0)$
$(-1, 1)$	$48 > 0$	$6 > 0$	$f$ a un minimum local en $(-1, 1)$
$(-1, -1)$	$48 > 0$	$6 > 0$	$f$ a un minimum local en $(-1, -1)$

### Exercice 3.11

Une montagne a la forme de la surface  $z(x, y) = 2xy - 2x^2 - y^2 - 8x + 6y + 4$  (l'unité de mesure est de 100 mètres). Si le niveau de la mer correspond à  $z = 0$ , quelle est la hauteur de la montagne ?

#### Correction

Il s'agit d'évaluer  $z(x, y)$  dans le point de maximum. Cherchons d'abord les points critiques :

$$\nabla z(x, y) = \begin{pmatrix} 2y - 4x - 8 \\ 2x - 2y + 6 \end{pmatrix}$$

et  $\nabla z(x, y) = \mathbf{0}$ ssi  $(x, y) = (-1, 2)$ . On établie la nature du point critique en étudiant le déterminant de la matrice hessienne :

$$\partial_{xx}f(x, y) = -4 < 0, \quad \partial_{yy}f(x, y) = -2, \quad \partial_{xy}f(x, y) = 2,$$

et  $\partial_{xx}f(-1, 2)\partial_{yy}f(-1, 2) - (\partial_{xy}f(-1, 2))^2 = 4 > 0$  donc  $(-1, 2)$  est un maximum. Comme  $z(-1, 2) = 14$ , la montagne est haute 1400 mètre.

### Exercice 3.12

Si  $f$  est une fonction continue d'une seule variable réelle et si  $f$  admet deux maxima sur un intervalle alors il existe un minimum compris entre les deux maxima. Le but de cet exercice est de montrer que ce résultat ne s'étend pas en deux dimensions.

Considérons la fonction  $f: \mathbb{R}^2 \rightarrow \mathbb{R}$  définie par  $f(x, y) = 4y^2 e^x - 2y^4 - e^{4x}$ . Montrer que cette fonction admet deux maxima mais aucun autre point critique.

**Correction**

\*  $f$  est de classe  $\mathcal{C}^2$  dans son domaine de définition, l'ouvert  $\mathbb{R}^2$ .

\* Recherche de points critiques :

$$\nabla f(x, y) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \begin{cases} 4y^2 e^x - 4e^{4x} = 0 \\ 8ye^x - 8y^3 = 0 \end{cases} \iff \begin{cases} 4e^x(y^2 - e^{3x}) = 0 \\ 8y(e^x - y^2) = 0 \end{cases} \begin{array}{l} \nearrow \\ \searrow \end{array} \begin{cases} y = 0 \\ -4e^{4x} = 0 \\ e^x = y^2 \\ y^2 = e^{3x} = y^6 \end{cases} \iff (x, y) = (0, \pm 1).$$

On a deux points critiques :  $(0, 1)$  et  $(0, -1)$ .

\* Nature des points critiques :

$$H_f(x, y) = \begin{pmatrix} 4y^2 e^x - 16e^{4x} & 8ye^x \\ 8ye^x & 8e^x - 24y^2 \end{pmatrix}, \quad \det(H_f(x, y)) = 32e^x((e^x - 3y^2)(y^2 - 4e^{3x}) - 2y^2 e^x).$$

$\det(H_f(0, \pm 1)) = 128 > 0$  et  $\partial_{xx}f(0, \pm 1) = -12 < 0$  donc les points  $(0, \pm 1)$  sont des maxima.

**Exercice 3.13**

Déterminer et établir la nature des points critiques des fonctions  $f: \mathbb{R}^2 \rightarrow \mathbb{R}$  définies par

- |  |   |  |
|--|---|--|
| 1. $f(x, y) = x^2 + xy + y^2 + y$                          | 2. $f(x, y) = xy - 2x - 2y - x^2 - y^2$                   | 3. $f(x, y) = (x - y)(1 - xy)$                             |
| 4. $f(x, y) = y^3 + 3x^2y - 6x^2 - 6y^2 + 2$               | 5. $f(x, y) = x^3 + y^3 - 3xy + 3$                        | 6. $f(x, y) = xy(1 - x - y)$                               |
| 7. $f(x, y) = x^3 - 12xy + 8y^3$                           | 8. $f(x, y) = xy + \frac{1}{x} + \frac{1}{y}$             | 9. $f(x, y) = e^x \cos(y)$                                 |
| 10. $f(x, y) = y \cos(x)$                                  | 11. $f(x, y) = y^2 + xy \ln(x)$                           | 12. $f(x, y) = \frac{x^2 y}{2} + x^2 + \frac{y^3}{3} - 4y$ |
| 13. $f(x, y) = \frac{x^2 y}{2} - x^2 + \frac{y^3}{3} - 4y$ | 14. $f(x, y) = \frac{xy^2}{2} + \frac{x^3}{3} - 4x + y^2$ | 15. $f(x, y) = (x^2 - y^2)e^{(-x^2 - y^2)}$                |
| 16. $f(x, y) = (y^2 - x^2)e^{(-x^2 - y^2)}$                | 17. $f(x, y) = x^4 + y^4 - 2(x - y)^2$                    | 18. $f(x, y) = x^4 + y^4 - 4(x - y)^2$                     |
| 19. $f(x, y, z) = \frac{x^2}{2} + xyz - z + y$             | 20. $f(x, y) = (x - 1)^2 + 2y^2$                          | 21. $f(x, y) = x^2 + xy + y^2 - 2x - y$                    |
| 22. $f(x, y) = x^3 y^2 (6 - x - y)$                        | 23. $f(x, y) = e^{x-y}(x^2 - 2y^2)$                       | 24. $f(x, y) = \frac{8}{x} + \frac{x}{y} + y$              |
| 25. $f(x, y) = x^2 - \cos(y)$                              | 26. $f(x, y) = (x^2 + y^2)e^{-(x^2 + y^2)}$               | 27. $f(x, y) = x^3 + y^2 - 6(x^2 - y^2)$                   |
| 28. $f(x, y) = (x^2 + y^2 - y^3)e^{-y}$                    |   |  |

**Correction**

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

\*  $f$  est de classe  $\mathcal{C}^2$  dans son domaine de définition, l'ouvert  $\mathbb{R}^2$ .

\* Recherche de points critiques :

$$\nabla f(x, y) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \begin{cases} 2x + y = 0 \\ x + 2y + 1 = 0 \end{cases} \iff (x, y) = \left(\frac{1}{3}, -\frac{2}{3}\right).$$

On a un unique point critique :  $(\frac{1}{3}, -\frac{2}{3})$ .

\* Nature du point critique :

$$H_f(x, y) = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \quad \det(H_f(x, y)) = 3.$$

$\det(H_f(\frac{1}{3}, -\frac{2}{3})) > 0$  et  $\partial_{xx}f(\frac{1}{3}, -\frac{2}{3}) > 0$  donc  $(\frac{1}{3}, -\frac{2}{3})$  est un minimum.

2.  $f(x, y) = xy - 2x - 2y - x^2 - y^2$

\*  $f$  est de classe  $\mathcal{C}^2$  dans son domaine de définition, l'ouvert  $\mathbb{R}^2$ .

\* Recherche de points critiques :

$$\nabla f(x, y) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \begin{cases} y - 2 - 2x = 0 \\ x - 2 - 2y = 0 \end{cases} \iff (x, y) = (-2, -2).$$

On a un unique point critique :  $(-2, -2)$ .

\* Nature du point critique :

$$H_f(x, y) = \begin{pmatrix} -2 & 1 \\ 1 & -2 \end{pmatrix}, \quad \det(H_f(x, y)) = 3.$$

$\det(H_f(-2, -2)) > 0$  et  $\partial_{xx}f(-2, -2) < 0$  donc  $(-2, -2)$  est un maximum.

3.  $f(x, y) = (x - y)(1 - xy)$

\*  $f$  est de classe  $\mathcal{C}^2$  dans son domaine de définition, l'ouvert  $\mathbb{R}^2$ .

\* Recherche de points critiques :

$$\nabla f(x, y) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \begin{cases} 1 - 2xy + y^2 = 0 \\ -1 - x^2 + 2xy = 0 \end{cases} \iff (x, y) \in \{(-1, -1), (1, 1)\}.$$

On a deux points critiques :  $(-1, -1)$  et  $(1, 1)$ .

\* Nature des points critiques :

$$H_f(x, y) = \begin{pmatrix} -2y & -2x+2y \\ -2x+2y & 2x \end{pmatrix}, \quad \det(H_f(x, y)) = -4xy - 4(y - x)^2.$$

$\det(H_f(-1, -1)) < 0$  donc  $(-1, -1)$  est un point-selle ;

$\det(H_f(1, 1)) < 0$  donc  $(1, 1)$  est un point-selle.

4.  $f(x, y) = y^3 + 3x^2y - 6x^2 - 6y^2 + 2$

\*  $f$  est de classe  $\mathcal{C}^2$  dans son domaine de définition, l'ouvert  $\mathbb{R}^2$ .

\* Recherche de points critiques :

$$\nabla f(x, y) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \begin{cases} 6xy - 12x = 0 \\ 3y^2 + 3x^2 - 12y = 0 \end{cases} \iff (x, y) \in \{(0, 0), (0, 4), (2, 2), (-2, 2)\}.$$

On a quatre points critiques :  $(0, 0)$ ,  $(0, 4)$ ,  $(2, 2)$  et  $(-2, 2)$ .

\* Nature des points critiques :

$$H_f(x, y) = \begin{pmatrix} 6y-12 & 6x \\ 6x & 6y-12 \end{pmatrix}, \quad \det(H_f(x, y)) = (6y - 12)^2 - 36x^2 = 36((y - 2)^2 - x^2).$$

$\det(H_f(0, 0)) > 0$  et  $\partial_{xx}f(0, 0) < 0$  donc  $(0, 0)$  est un maximum ;

$\det(H_f(0, 4)) > 0$  et  $\partial_{xx}f(0, 4) > 0$  donc  $(0, 4)$  est un minimum ;

$\det(H_f(2, 2)) < 0$  donc  $(2, 2)$  est un point-selle ;

$\det(H_f(-2, 2)) < 0$  donc  $(-2, 2)$  est un point-selle.

5.  $f(x, y) = x^3 + y^3 - 3xy + 3$ .

\*  $f$  est de classe  $\mathcal{C}^2$  dans son domaine de définition, l'ouvert  $\mathbb{R}^2$ .

\* Recherche de points critiques :

$$\nabla f(x, y) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \begin{cases} 3(x^2 - y) = 0 \\ 3(y^2 - x) = 0 \end{cases} \iff (x, y) \in \{(0, 0), (1, 1)\}.$$

On a deux points critiques :  $(0, 0)$  et  $(1, 1)$ .

\* Nature des points critiques :

$$H_f(x, y) = \begin{pmatrix} 6x & -3 \\ -3 & 6y \end{pmatrix}, \quad \det(H_f(x, y)) = 36xy - 9.$$

$\det(H_f(0, 0)) = -9 < 0$  donc  $(0, 0)$  est un point-selle ;

$\det(H_f(1, 1)) = 27 > 0$  et  $\partial_{xx}f(1, 1) = 6 > 0$ , donc  $(1, 1)$  est un minimum.

6.  $f(x, y) = xy(1 - x - y)$

\*  $f$  est de classe  $\mathcal{C}^2$  dans son domaine de définition, l'ouvert  $\mathbb{R}^2$ .

\* Recherche de points critiques :

$$\nabla f(x, y) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \begin{cases} y - 2xy - y^2 = 0 \\ x - x^2 - 2xy = 0 \end{cases} \iff (x, y) \in \{(0, 0), (1, 0), (0, 1), \left(\frac{1}{3}, \frac{1}{3}\right)\}.$$

On a quatre points critiques :  $(0, 0)$ ,  $(1, 0)$ ,  $(0, 1)$  et  $(\frac{1}{3}, \frac{1}{3})$ .

- \* Nature des points critiques :

$$H_f(x, y) = \begin{pmatrix} -2y & 1-2x-2y \\ 1-2x-2y & -2x \end{pmatrix}, \quad \det(H_f(x, y)) = 4xy - (1-2x-2y)^2$$

$\det(H_f(0, 0)) < 0$  donc  $(0, 0)$  est un point-selle ;

$\det(H_f(1, 0)) < 0$  donc  $(1, 0)$  est un point-selle ;

$\det(H_f(0, 1)) < 0$  donc  $(0, 1)$  est un point-selle ;

$\det(H_f(\frac{1}{3}, \frac{1}{3})) > 0$  et  $\partial_{xx}f(\frac{1}{3}, \frac{1}{3}) < 0$  donc  $(\frac{1}{3}, \frac{1}{3}) < 0$  est un maximum.

7.  $f(x, y) = x^3 - 12xy + 8y^3$

- \*  $f$  est de classe  $\mathcal{C}^2$  dans son domaine de définition, l'ouvert  $\mathbb{R}^2$ .

- \* Recherche de points critiques :

$$\nabla f(x, y) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \begin{cases} 3x^2 - 12y = 0 \\ -12x + 24y^2 = 0 \end{cases} \iff (x, y) \in \{(0, 0), (2, 1)\}.$$

On a deux points critiques :  $(0, 0)$ , et  $(2, 1)$ .

- \* Nature des points critiques :

$$H_f(x, y) = \begin{pmatrix} 6x & -12 \\ -12 & 48y \end{pmatrix}, \quad \det(H_f(x, y)) = 144(2xy - 1)$$

$\det(H_f(0, 0)) < 0$  donc  $(0, 0)$  est un point-selle ;

$\det(H_f(2, 1)) > 0$  et  $\partial_{xx}f(2, 1) > 0$  donc  $(2, 1)$  est un minimum.

8.  $f(x, y) = xy + \frac{1}{x} + \frac{1}{y}$

- \*  $f$  est de classe  $\mathcal{C}^2$  dans son domaine de définition, l'ouvert  $\mathbb{R}^2 \setminus \{(0, \kappa) \mid \kappa \in \mathbb{R}\} \setminus \{(\kappa, 0) \mid \kappa \in \mathbb{R}\}$ .

- \* Recherche de points critiques :

$$\nabla f(x, y) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \begin{cases} y - \frac{1}{x^2} = 0 \\ x - \frac{1}{y^2} = 0 \end{cases} \iff (x, y) = (1, 1).$$

On a un unique point critique :  $(1, 1)$ .

- \* Nature du point critique :

$$H_f(x, y) = \begin{pmatrix} \frac{2}{x^3} & 1 \\ 1 & \frac{2}{y^3} \end{pmatrix}, \quad \det(H_f(x, y)) = \frac{4}{(xy)^3} - 1$$

$\det(H_f(1, 1)) > 0$  et  $\partial_{xx}f(1, 1) > 0$  donc  $(1, 1)$  est un minimum.

9.  $f(x, y) = e^x \cos(y)$

- \*  $f$  est de classe  $\mathcal{C}^2$  dans son domaine de définition, l'ouvert  $\mathbb{R}^2$ .

- \* Recherche de points critiques :

$$\nabla f(x, y) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \begin{cases} e^x \cos(y) = 0 \\ -e^x \sin(y) = 0 \end{cases} \iff \exists (x, y) \in \mathbb{R}^2.$$

Cette fonction n'admet aucun point critique.

10.  $f(x, y) = y \cos(x)$

- \*  $f$  est de classe  $\mathcal{C}^2$  dans son domaine de définition, l'ouvert  $\mathbb{R}^2$ .

- \* Recherche de points critiques :

$$\nabla f(x, y) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \begin{cases} -y \sin(x) = 0 \\ \cos(x) = 0 \end{cases} \iff (x, y) = \left(\frac{\pi}{2} + \kappa\pi, 0\right), \quad \kappa \in \mathbb{Z}.$$

On a une infinité de points critiques alignés sur la droite d'équation  $y = 0$  et qui ont ordonnée  $x = \frac{\pi}{2} + \kappa\pi$  avec  $\kappa \in \mathbb{Z}$ .

\* Nature des points critiques :

$$H_f(x, y) = \begin{pmatrix} -y \cos(x) & -\sin(x) \\ -\sin(x) & 0 \end{pmatrix}, \quad \det(H_f(x, y)) = -\sin^2(x)$$

$\det(H_f(\frac{\pi}{2} + \kappa\pi, 0)) < 0$  donc ils sont tous des points-selle.

11.  $f(x, y) = y^2 + xy \ln(x)$

\*  $f$  est de classe  $\mathcal{C}^2$  dans son domaine de définition, l'ouvert  $\mathcal{D} = \{(x, y) \in \mathbb{R}^2 \mid x > 0\}$ .

\* Recherche de points critiques :

$$\nabla f(x, y) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \begin{cases} y(\ln(x) + 1) = 0 \\ 2y + x \ln(x) = 0 \end{cases} \iff (x, y) \in \left\{ (1, 0); \left(\frac{1}{e}, \frac{1}{2e}\right) \right\}$$

On a deux points critiques :  $(1, 0)$  et  $(\frac{1}{e}, \frac{1}{2e})$ .

\* Nature des points critiques :

$$H_f(x, y) = \begin{pmatrix} \frac{y}{x} & 1 + \ln(x) \\ 1 + \ln(x) & 2 \end{pmatrix}, \quad \det(H_f(x, y)) = 2\frac{y}{x} - (1 + \ln(x))^2$$

$\det(H_f(1, 0)) < 0$  donc  $(1, 0)$  est un point-selle;

$\det(H_f(\frac{1}{e}, \frac{1}{2e})) > 0$  et  $\partial_{xx}f(\frac{1}{e}, \frac{1}{2e}) > 0$  donc  $(\frac{1}{e}, \frac{1}{2e})$  est un minimum.

12.  $f(x, y) = \frac{x^2 y}{2} + x^2 + \frac{y^3}{3} - 4y$

\*  $f$  est de classe  $\mathcal{C}^2$  dans son domaine de définition, l'ouvert  $\mathbb{R}^2$ .

\* Recherche de points critiques :

$$\nabla f(x, y) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \begin{cases} xy + 2x = 0 \\ \frac{x^2}{2} + y^2 - 4 = 0 \end{cases} \iff (x, y) \in \{(0, -2), (0, 2)\}.$$

On a deux points critiques :  $(0, -2)$  et  $(0, 2)$ .

\* Nature des points critiques :

$$H_f(x, y) = \begin{pmatrix} y+2 & x \\ x & 2y \end{pmatrix}, \quad \det(H_f(x, y)) = 2y(y+2) - x$$

$\det(H_f(0, 2)) > 0$  et  $\partial_{xx}f(0, 2) = 4 > 0$  donc  $(0, 2)$  est un minimum pour  $f$  ;

comme  $\det(H_f(0, -2)) = 0$ , on ne peut pas conclure en utilisant la matrice hessienne (l'étude du signe de la distance dans ce cas est trop compliquée).

13.  $f(x, y) = \frac{x^2 y}{2} - x^2 + \frac{y^3}{3} - 4y$

\*  $f$  est de classe  $\mathcal{C}^2$  dans son domaine de définition, l'ouvert  $\mathbb{R}^2$ .

\* Recherche de points critiques :

$$\nabla f(x, y) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \begin{cases} xy - 2x = 0 \\ \frac{x^2}{2} + y^2 - 4 = 0 \end{cases} \iff (x, y) \in \{(0, -2), (0, 2)\}.$$

On a deux points critiques :  $(0, -2)$  et  $(0, 2)$ .

\* Nature des points critiques :

$$H_f(x, y) = \begin{pmatrix} y-2 & x \\ x & 2y \end{pmatrix}, \quad \det(H_f(x, y)) = 2y(y-2) - x$$

$\det(H_f(0, -2)) > 0$  et  $\partial_{xx}f(0, -2) < 0$  donc  $(0, -2)$  est un maximum pour  $f$  ;

comme  $\det(H_f(0, 2)) = 0$ , on ne peut pas conclure en utilisant la matrice hessienne (l'étude du signe de la distance dans ce cas est trop compliquée).

14.  $f(x, y) = \frac{xy^2}{2} + \frac{x^3}{3} - 4x + y^2$

\*  $f$  est de classe  $\mathcal{C}^2$  dans son domaine de définition, l'ouvert  $\mathbb{R}^2$ .

- ★ Recherche de points critiques :

$$\nabla f(x, y) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \begin{cases} \frac{y^2}{2} + x^2 - 4 = 0 \\ xy + 2y = 0 \end{cases} \iff (x, y) \in \{(-2, 0), (2, 0)\}.$$

On a deux points critiques :  $(0, -2)$  et  $(0, 2)$ .

- ★ Nature des points critiques :

$$H_f(x, y) = \begin{pmatrix} 2x & y \\ y & x+2 \end{pmatrix}, \quad \det(H_f(x, y)) = 2x(x+2) - y$$

$\det(H_f(2, 0)) > 0$  et  $\partial_{xx}f(2, 0) = 4 > 0$  donc  $(2, 0)$  est un minimum pour  $f$  ;

comme  $\det(H_f(-2, 0)) = 0$ , on ne peut pas conclure en utilisant la matrice hessienne (l'étude du signe de la distance dans ce cas est trop compliquée).

15.  $f(x, y) = (x^2 - y^2)e^{(-x^2 - y^2)}$

- ★  $f$  est de classe  $\mathcal{C}^2$  dans son domaine de définition, l'ouvert  $\mathbb{R}^2$ .

- ★ Recherche de points critiques :

$$\nabla f(x, y) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \begin{cases} 2x(1 - x^2 + y^2)e^{(-x^2 - y^2)} = 0 \\ 2y(-1 - x^2 + y^2)e^{(-x^2 - y^2)} = 0 \end{cases} \iff (x, y) \in \{(0, 0), (0, 1), (0, -1), (1, 0), (-1, 0)\}.$$

On a 5 points critiques :  $(0, 0)$ ,  $(0, 1)$ ,  $(0, -1)$ ,  $(1, 0)$  et  $(-1, 0)$ .

- ★ Nature des points critiques :

$$\begin{aligned} \partial_{xx}f(x, y) &= 2e^{(-x^2 - y^2)}(1 - 5x^2 + y^2 + 2x^4 - 2x^2y^2), \\ \partial_{xy}f(x, y) &= 4xy(x^2 - y^2)e^{(-x^2 - y^2)}, \\ \partial_{yy}f(x, y) &= 2e^{(-x^2 - y^2)}(-1 - x^2 + 5y^2 + 2x^2y^2 - 2y^4). \end{aligned}$$

$$H_f(x, y) = \begin{pmatrix} \partial_{xx}f(x, y) & \partial_{xy}f(x, y) \\ \partial_{xy}f(x, y) & \partial_{yy}f(x, y) \end{pmatrix}, \quad \det(H_f(x, y)) = \partial_{xx}f(x, y)\partial_{yy}f(x, y) - (\partial_{xy}f(x, y))^2$$

On a alors

$(x_0, y_0)$	$\partial_{xx}f(x_0, y_0)$	$\partial_{xy}f(x_0, y_0)$	$\partial_{yy}f(x_0, y_0)$	$\det(H_f(x_0, y_0))$	
$(0, 0)$	2	0	-2	-4	c'est un point-selle
$(1, 0)$	$-\frac{4}{e}$	0	$-\frac{4}{e}$	$\frac{16}{e^2}$	c'est un maximum
$(-1, 0)$	$-\frac{4}{e}$	0	$-\frac{4}{e}$	$\frac{16}{e^2}$	c'est un maximum
$(0, 1)$	$\frac{4}{e}$	0	$\frac{4}{e}$	$\frac{16}{e^2}$	c'est un minimum
$(0, -1)$	$\frac{4}{e}$	0	$\frac{4}{e}$	$\frac{16}{e^2}$	c'est un minimum

16.  $f(x, y) = (y^2 - x^2)e^{(-x^2 - y^2)}$

- ★  $f$  est de classe  $\mathcal{C}^2$  dans son domaine de définition, l'ouvert  $\mathbb{R}^2$ .

- ★ Recherche de points critiques :

$$\nabla f(x, y) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \begin{cases} 2x(-1 + x^2 - y^2)e^{(-x^2 - y^2)} = 0 \\ 2y(1 + x^2 - y^2)e^{(-x^2 - y^2)} = 0 \end{cases} \iff (x, y) \in \{(0, 0), (0, 1), (0, -1), (1, 0), (-1, 0)\}.$$

On a 5 points critiques :  $(0, 0)$ ,  $(0, 1)$ ,  $(0, -1)$ ,  $(1, 0)$  et  $(-1, 0)$ .

- ★ Nature des points critiques :

$$\begin{aligned} \partial_{xx}f(x, y) &= -2e^{(-x^2 - y^2)}(1 - 5x^2 + y^2 + 2x^4 - 2x^2y^2), \\ \partial_{xy}f(x, y) &= -4xy(x^2 - y^2)e^{(-x^2 - y^2)}, \\ \partial_{yy}f(x, y) &= -2e^{(-x^2 - y^2)}(-1 - x^2 + 5y^2 + 2x^2y^2 - 2y^4). \end{aligned}$$

$$H_f(x, y) = \begin{pmatrix} \partial_{xx}f(x, y) & \partial_{xy}f(x, y) \\ \partial_{xy}f(x, y) & \partial_{yy}f(x, y) \end{pmatrix}, \quad \det(H_f(x, y)) = \partial_{xx}f(x, y)\partial_{yy}f(x, y) - (\partial_{xy}f(x, y))^2$$

On a alors

$(x_0, y_0)$	$\partial_{xx}f(x_0, y_0)$	$\partial_{xy}f(x_0, y_0)$	$\partial_{yy}f(x_0, y_0)$	$\det(H_f(x_0, y_0))$	
$(0, 0)$	-2	0	2	-4	c'est un point-selle
$(1, 0)$	$\frac{4}{e}$	0	$\frac{4}{e}$	$\frac{16}{e^2}$	c'est un minimum
$(-1, 0)$	$\frac{4}{e}$	0	$\frac{4}{e}$	$\frac{16}{e^2}$	c'est un minimum
$(0, 1)$	$-\frac{4}{e}$	0	$-\frac{4}{e}$	$\frac{16}{e^2}$	c'est un maximum
$(0, -1)$	$-\frac{4}{e}$	0	$-\frac{4}{e}$	$\frac{16}{e^2}$	c'est un maximum

17.  $f(x, y) = x^4 + y^4 - 2(x - y)^2$

- \*  $f$  est de classe  $C^2$  dans son domaine de définition, l'ouvert  $\mathbb{R}^2$ . Comme la restriction  $f(x, 0) = x^4 - 2x^2$  tend vers  $+\infty$  pour  $x$  qui tend vers  $\pm\infty$ , il n'y a pas de maximum global sur  $\mathbb{R}^2$ . Comme  $\mathbb{R}^2$  est ouvert, un extrémum relatif de  $f$  vérifie la condition nécessaire  $\nabla f(x, y) = 0$ .

- \* Recherche de points critiques :

$$\nabla f(x, y) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \begin{cases} 4(x^3 - x + y) = 0 \\ 4(y^3 + x - y) = 0 \end{cases} \iff (x, y) \in \{(0, 0), (\sqrt{2}, -\sqrt{2}), (-\sqrt{2}, \sqrt{2})\}.$$

On a 3 points critiques : <sup>1</sup>  $(0, 0)$ ,  $(\sqrt{2}, -\sqrt{2})$  et  $(-\sqrt{2}, \sqrt{2})$  (on note que  $f(x, y) = f(-x, -y)$ ).

- \* Nature des points critiques :

$$H_f(x, y) = \begin{pmatrix} 12x^2 - 4 & 4 \\ 4 & 12y^2 - 4 \end{pmatrix}, \quad \det(H_f(x, y)) = 16((3x^2 - 1)(3y^2 - 1) - 1).$$

$\det(H_f(\sqrt{2}, -\sqrt{2})) = 384 > 0$  et  $\partial_{xx}f(\sqrt{2}, -\sqrt{2}) = 20 > 0$  donc  $(\sqrt{2}, -\sqrt{2})$  est un minimum pour  $f$  ;  
 $\det(H_f(-\sqrt{2}, \sqrt{2})) = 384 > 0$  et  $\partial_{xx}f(-\sqrt{2}, \sqrt{2}) = 20 > 0$  donc  $(-\sqrt{2}, \sqrt{2})$  est un minimum pour  $f$  ;  
comme  $\det(H_f(0, 0)) = 0$ , on ne peut pas conclure en utilisant la matrice hessienne.

Pour connaître la nature du point  $(0, 0)$  on étudie le signe de  $d(h, k) = f(h, k) - f(0, 0)$  pour  $h$  et  $k$  voisins de 0 :

$$d(h, k) = h^4 + k^4 - 2(h - k)^2;$$

comme  $d(h, 0) = (h^2 - 2)h^2 < 0$  lorsque  $h$  est voisin de 0 mais  $d(h, h) = 2h^4 > 0$ , alors  $(0, 0)$  est un point-selle.

Remarquons qu'avec des transformations algébriques, on peut réécrire la fonction sous la forme

$$f(x, y) = (x^2 - 2)^2 + (y^2 - 2)^2 + 2(x + y)^2 - 8 \geq 8 \quad \forall (x, y) \in \mathbb{R}^2.$$

Comme  $f(\sqrt{2}, -\sqrt{2}) = f(-\sqrt{2}, \sqrt{2}) = -8$ , les points  $(\sqrt{2}, -\sqrt{2})$  et  $(-\sqrt{2}, \sqrt{2})$  sont des minima globaux.

18.  $f(x, y) = x^4 + y^4 - 4(x - y)^2$

- \*  $f$  est de classe  $C^2$  dans son domaine de définition, l'ouvert  $\mathbb{R}^2$ . Comme la restriction  $f(x, 0) = x^4 - 4x^2$  tend vers  $+\infty$  pour  $x$  qui tend vers  $\pm\infty$ , il n'y a pas de maximum global sur  $\mathbb{R}^2$ . Comme  $\mathbb{R}^2$  est ouvert, un extrémum relatif de  $f$  vérifie la condition nécessaire  $\nabla f(x, y) = 0$ .

- \* Recherche de points critiques :

$$\nabla f(x, y) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \begin{cases} 4(x^3 - 2x + 2y) = 0 \\ 4(y^3 + 2x - 2y) = 0 \end{cases} \iff (x, y) \in \{(0, 0), (2, -2), (-2, 2)\}.$$

On a 3 points critiques : <sup>2</sup>  $(0, 0)$ ,  $(2, -2)$  et  $(-2, 2)$  (on note que  $f(x, y) = f(-x, -y)$ ).

---


$$\begin{aligned} 1. \quad & \begin{cases} 4x^3 - 4x + 4y = 0 \\ 4y^3 + 4x - 4y = 0 \end{cases} \implies \begin{cases} x^3 + y^3 = 0 \\ y^3 + x - y = 0 \end{cases} \implies \begin{cases} x = -y \\ (y^2 - 2)y = 0 \end{cases} \\ 2. \quad & \begin{cases} 4x^3 - 8x + 8y = 0 \\ 4y^3 + 8x - 8y = 0 \end{cases} \implies \begin{cases} x^3 - 2(x - y) = 0 \\ y^3 + 2(x - y) = 0 \end{cases} \implies \begin{cases} x^3 + y^3 = 0 \\ y^3 + 2x - 2y = 0 \end{cases} \implies \begin{cases} x = -y \\ (y^2 - 4)y = 0 \end{cases} \implies \begin{cases} x = -y \\ y = 0 \text{ ou } y = 2 \text{ ou } y = -2 \end{cases} \end{aligned}$$

- \* Nature des points critiques :

$$H_f(x, y) = \begin{pmatrix} 12x^2 - 8 & 8 \\ 8 & 12y^2 - 8 \end{pmatrix}, \quad \det(H_f(x, y)) = 48(3x^2y^2 - 2(x^2 + y^2)).$$

- \*  $\det(H_f(2, -2)) = 1536 > 0$  et  $\partial_{xx}f(2, -2) = 40 > 0$  donc  $(2, -2)$  est un minimum local pour  $f$  ;
- \*  $\det(H_f(-2, 2)) = 1536 > 0$  et  $\partial_{xx}f(-2, 2) = 40 > 0$  donc  $(-2, 2)$  est un minimum local pour  $f$  ;
- \* comme  $\det(H_f(0, 0)) = 0$ , on ne peut pas conclure en utilisant la matrice hessienne. Pour connaître la nature du point  $(0, 0)$  on étudie le signe de  $d(h, k) = f(h, k) - f(0, 0)$  pour  $h$  et  $k$  voisins de 0 :

$$d(h, k) = h^4 + k^4 - 4(h - k)^2;$$

comme  $d(h, 0) = (h^2 - 4)h^2 < 0$  lorsque  $h$  est voisin de 0 mais  $d(h, h) = 2h^4 > 0$ , alors  $(0, 0)$  est un point-selle.

Remarquons qu'avec des transformations algébriques, on peut réécrire la fonction sous la forme

$$f(x, y) = (x^2 - 4)^2 + (y^2 - 4)^2 + 4(x + y)^2 - 32 \geq -32 \quad \forall (x, y) \in \mathbb{R}^2.$$

Comme  $f(2, -2) = f(-2, 2) = -32$ , les points  $(2, -2)$  et  $(-2, 2)$  sont des minima globaux.

19.  $f(x, y, z) = \frac{x^2}{2} + xyz - z + y$

- \*  $f$  est définie sur  $\mathbb{R}^3$  à valeur dans  $\mathbb{R}$  ; comme la restriction  $f(0, 0, z) = -z$  tend vers  $\pm\infty$  pour  $z$  qui tend vers  $\mp\infty$ , il n'y a pas d'extremum global sur  $\mathbb{R}^3$ . Comme  $\mathbb{R}^3$  est ouvert, un extrémum relatif de  $f$  vérifie la condition nécessaire  $\nabla f(x, y, z) = 0$ .

- \* Recherche de points critiques :

$$\nabla f(x, y, z) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \begin{cases} x + yz = 0 \\ xz + 1 = 0 \\ xy - 1 = 0 \end{cases} \iff (x, y, z) = (1, 1, -1).$$

Il n'y a qu'un point critique :  $(1, 1, -1)$ .

- \* Nature du point critique : on étudie le signe de  $\Delta f(h, k, l) \equiv f(1+h, 1+k, -1+l)$  pour  $h, k$  et  $l$  voisins de 0 (les termes de degré 1 en  $h, k$  et  $l$  doivent disparaître) :

$$\Delta f(h, k, l) = \frac{h^2 + 1 + 2h}{2} + (1+h)(1+k)(-1+l) - (-1+l) + (1+k) - \frac{3}{2} = \frac{h^2}{2} + hkl + hl - hk + kl.$$

Il ne reste que transformer  $\Delta f$  si on pense qu'il s'agit d'un extrémum ou fournir des restrictions qui se contredisent si on pense que ce n'est pas un extrémum. Comme les deux restrictions à deux courbes continues passant par l'origine  $\Delta f(h, 0, h) = \frac{3}{2}h^2 > 0$  et  $\Delta f(h, h, 0) = -\frac{1}{2}h^2 < 0$  donnent des signes différents, on conclut que ce n'est pas un extrémum.

20.  $f(x, y) = (x - 1)^2 + 2y^2$

- \*  $f$  est de classe  $\mathcal{C}^2$  dans son domaine de définition, l'ouvert  $\mathbb{R}^2$ .

- \* Recherche de points critiques :

$$\nabla f(x, y) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \begin{cases} 2x - 2 = 0 \\ 4y = 0 \end{cases} \iff (x, y) = (1, 0).$$

On a un seul point critique :  $(1, 0)$ .

- \* Nature du point critique :

$$H_f(x, y) = \begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix}, \quad \det(H_f(x, y)) = 8.$$

$\det(H_f(1, 0)) = 8 > 0$  et  $\partial_{xx}f(1, 0) = 2 > 0$  donc  $(1, 0)$  est un minimum pour  $f$ .

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

- \*  $f$  est de classe  $\mathcal{C}^2$  dans son domaine de définition, l'ouvert  $\mathbb{R}^2$ .

- \* Recherche de points critiques :

$$\nabla f(x, y) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \begin{cases} 2x + y - 2 = 0 \\ x + 2y - 1 = 0 \end{cases} \iff (x, y) = (1, 0).$$

On a un seul point critique :  $(1, 0)$ .

- \* Nature du point critique :

$$H_f(x, y) = \begin{pmatrix} 2 & 1 \\ 1 & 4 \end{pmatrix}, \quad \det(H_f(x, y)) = 7.$$

$\det(H_f(1, 0)) = 8 > 0$  et  $\partial_{xx}f(1, 0) = 2 > 0$  donc  $(1, 0)$  est un minimum pour  $f$ .

22.  $f(x, y) = x^3y^2(6 - x - y)$

- \*  $f$  est de classe  $\mathcal{C}^2$  dans son domaine de définition, l'ouvert  $\mathbb{R}^2$ .

- \* Recherche de points critiques :

$$\nabla f(x, y) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \begin{cases} 3x^2y^2(6 - x - y) - x^3y^2 = 0 \\ 2x^3y(6 - x - y) - x^3y^2 = 0 \end{cases} \iff (x, y) \in \{(3, 2), (t, 0), (0, t) \mid t \in \mathbb{R}\}.$$

On a une infinité de points critiques : les points  $(t, 0)$  et  $(0, t)$  pour  $t \in \mathbb{R}$  sont des points critiques ainsi que le point  $(3, 2)$ .

- \* Nature des points critiques :

$$\begin{aligned} \partial_{xx}f(x, y) &= 6xy^2(6 - x - y) - 6x^2y^2, \\ \partial_{xy}f(x, y) &= 6x^2y(6 - x - y) - 3x^2y^2 - 2x^3y, \\ \partial_{yy}f(x, y) &= 2x^3(6 - x - y) - 4x^3y. \end{aligned}$$

$$H_f(x, y) = \begin{pmatrix} \partial_{xx}f(x, y) & \partial_{xy}f(x, y) \\ \partial_{xy}f(x, y) & \partial_{yy}f(x, y) \end{pmatrix}, \quad \det(H_f(x, y)) = \partial_{xx}f(x, y)\partial_{yy}f(x, y) - (\partial_{xy}f(x, y))^2$$

$\det(H_f(3, 2)) > 0$  et  $\partial_{xx}f(3, 2) < 0$  donc  $(3, 2)$  est un maximum pour  $f$ .

$\det(H_f(t, 0)) = 0$  pour tout  $t \in \mathbb{R}$  : l'étude de la matrice hessienne ne permet pas de conclure pour les points sur l'axe d'équation  $y = 0$ . Pour connaître la nature de ces points on étudie le signe de  $d(h, k) = f(t+h, 0+k) - f(t, 0) = (t+h)^3k^2(6 - t - h - k)$  pour  $h$  et  $k$  proches de 0. On conclut que les points  $(t, 0)$  pour  $t < 0$  ou  $t > 6$  sont des maxima, les points  $(t, 0)$  pour  $0 < t < 6$  sont des minima et les points  $(0, 0)$  et  $(6, 0)$  sont des points-selle.

$\det(H_f(0, t)) = 0$  pour tout  $t \in \mathbb{R}$  : l'étude de la matrice hessienne ne permet pas de conclure pour les points sur les axes. Pour connaître la nature de ces points on étudie le signe de  $d(h, k) = f(0+h, t+k) - f(0, t) = h^3(t+k)^2(6 - t - h - k)$  pour  $h$  et  $k$  proches de 0. On conclut que les points  $(0, t)$  sont des points-selle pour tout  $t \in \mathbb{R}$ .

23.  $f(x, y) = e^{x-y}(x^2 - 2y^2)$

- \*  $f$  est de classe  $\mathcal{C}^2$  dans son domaine de définition, l'ouvert  $\mathbb{R}^2$ .

- \* Recherche de points critiques :

$$\nabla f(x, y) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \begin{cases} (x^2 - 2y^2 + 2x)e^{x-y} = 0 \\ (-x^2 + 2y^2 - 4y)e^{x-y} = 0 \end{cases} \iff (x, y) \in \{(0, 0), (-4, -2)\}.$$

On a deux points critiques :  $(0, 0)$  et  $(-4, -2)$ .

- \* Nature des points critiques :

$$\partial_{xx}f(x, y) = e^{x-y}(x^2 - 2y^2 + 4x + 2), \quad \partial_{xy}f(x, y) = e^{x-y}(-x^2 + 2y^2 - 2x - 4y), \quad \partial_{yy}f(x, y) = e^{x-y}(x^2 - 2y^2 + 8y - 4);$$

$$H_f(x, y) = \begin{pmatrix} \partial_{xx}f(x, y) & \partial_{xy}f(x, y) \\ \partial_{xy}f(x, y) & \partial_{yy}f(x, y) \end{pmatrix}, \quad \det(H_f(x, y)) = \partial_{xx}f(x, y)\partial_{yy}f(x, y) - (\partial_{xy}f(x, y))^2.$$

On en déduit que

$(x_0, y_0)$	$\partial_{xx}f(x_0, y_0)$	$\partial_{xy}f(x_0, y_0)$	$\partial_{yy}f(x_0, y_0)$	$\det(H_f(x_0, y_0))$	
$(-4, -2)$	$-6e^{-2}$	$8e^{-2}$	$-12e^{-2}$	$8e^{-4}$	maximum
$(0, 0)$	2	0	-4	-8	point-selle

24.  $f(x, y) = \frac{8}{x} + \frac{x}{y} + y$

- \*  $f$  est de classe  $\mathcal{C}^2$  dans son domaine de définition, l'ouvert  $\mathbb{R}^2 \setminus \{(x, y) \mid xy = 0\}$ .

- \* Recherche de points critiques :

$$\nabla f(x, y) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \begin{cases} \frac{1}{y} - \frac{8}{x^2} = 0 \\ 1 - \frac{x}{y^2} = 0 \end{cases} \iff (x, y) = (4, 2).$$

On a un unique point critique :  $(4, 2)$ .

- \* Nature du point critique :

$$H_f(x, y) = \begin{pmatrix} \frac{16}{x^3} & -\frac{1}{y^2} \\ -\frac{1}{y^2} & \frac{2x}{y^3} \end{pmatrix}, \quad \det(H_f(x, y)) = \frac{1}{y^3} \left( \frac{16}{x^2} - \frac{1}{y} \right).$$

$\det(H_f(4, 2)) > 0$  et  $\partial_{xx}f(4, 2) > 0$  donc  $(4, 2)$  est un minimum pour  $f$ .

25.  $f(x, y) = x^2 - \cos(y)$

- \*  $f$  est de classe  $\mathcal{C}^2$  dans son domaine de définition, l'ouvert  $\mathbb{R}^2$ .

- \* Recherche de points critiques :

$$\nabla f(x, y) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \begin{cases} 2x = 0 \\ \sin(y) = 0 \end{cases} \iff (x, y) \in \{(0, \kappa\pi) \mid \kappa \in \mathbb{Z}\}.$$

On a une infinité de points critiques qui s'écrivent  $(0, \kappa\pi)$  avec  $\kappa \in \mathbb{Z}$ .

- \* Nature du point critique :

$$H_f(x, y) = \begin{pmatrix} 2 & 0 \\ 0 & \cos(y) \end{pmatrix}, \quad \det(H_f(x, y)) = 2\cos(y).$$

$\det(H_f(0, \kappa\pi)) = (-1)^\kappa$  et  $\partial_{xx}f(0, \kappa\pi) > 0$  pour tout  $\kappa \in \mathbb{Z}$  donc  $(0, \kappa\pi)$  est un minimum si  $\kappa$  est pair et un point-selle si  $\kappa$  est impair.

26.  $f(x, y) = (x^2 + y^2)e^{-(x^2+y^2)}$ .

On peut remarquer que si on passe aux coordonnées polaires on obtient  $w(r) \equiv f(r \cos(\vartheta), r \sin(\vartheta)) = r^2 e^{-r^2}$ , autrement dit on obtient une fonction de la seule variable  $r > 0$  et on a  $w'(r) = 2r(1 - r^2)e^{-r^2}$  qui s'annule pour  $r = 1$  et dont l'étude des variations montre qu'il s'agit d'un minimum. Il faut étudier séparément le cas  $(x = 0, y = 0)$  car il n'est pas pris en compte lorsqu'on passe aux coordonnées polaires. Si on n'a pas remarqué cette symétrie, on étudie la fonction comme dans les cas précédents :

- \*  $f$  est de classe  $\mathcal{C}^2$  dans son domaine de définition, l'ouvert  $\mathbb{R}^2$ .

- \* Recherche de points critiques :

$$\nabla f(x, y) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \begin{cases} 2x(1 - x^2 - y^2)e^{-(x^2+y^2)} = 0 \\ 2y(1 - x^2 - y^2)e^{-(x^2+y^2)} = 0 \end{cases}$$

On a une infinité de points critiques : le point  $(0, 0)$  et les points  $(x, y)$  qui appartiennent au cercle  $x^2 + y^2 = 1$ .

- \* Nature du point critique : comme  $f(x, y) \geq 0$  pour tout  $(x, y) \in \mathbb{R}^2$  et  $f(x, y) = 0$  ssi  $(x, y) \neq (0, 0)$  ou  $(x, y)$  est tel que  $x^2 + y^2 - 1 = 0$ , on en déduit qu'ils sont des minima (le calcul des dérivées seconde porte à des calculs très longues et inutiles dans ce cas).

27.  $f(x, y) = x^3 + y^2 - 6(x^2 - y^2)$

- \*  $f$  est de classe  $\mathcal{C}^2$  dans son domaine de définition, l'ouvert  $\mathbb{R}^2$ .

- \* Recherche de points critiques :

$$\nabla f(x, y) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \begin{cases} 3x(x - 4) = 0 \\ 3y(y + 4) = 0 \end{cases} \iff (x, y) \in \{(0, 0), (0, -4), (4, 0), (4, -4)\}.$$

On a quatre points critiques :  $(0, 0)$ ,  $(0, -4)$ ,  $(4, 0)$  et  $(4, -4)$ .

- \* Nature des points critiques :

$$H_f(x, y) = \begin{pmatrix} 6(x-2) & 0 \\ 0 & 6(y+2) \end{pmatrix}, \quad \det(H_f(x, y)) = 36(x-2)(y+2).$$

$\det(H_f(0, 0)) < 0$  donc  $(0, 0)$  est un point-selle ;

$\det(H_f(0, -4)) > 0$  et  $\partial_{xx}f(0, -4) < 0$  donc  $(0, -4)$  est un maximum ;  
 $\det(H_f(4, 0)) > 0$  et  $\partial_{xx}f(4, 0) > 0$  donc  $(4, 0)$  est un minimum ;  
 $\det(H_f(4, -4)) < 0$   $(4, -4)$  est un point-selle.

28.  $f(x, y) = (x^2 + y^2 - y^3)e^{-y}$

- \*  $f$  est de classe  $\mathcal{C}^2$  dans son domaine de définition, l'ouvert  $\mathbb{R}^2$ .

- \* Recherche de points critiques :

$$\nabla f(x, y) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \begin{cases} 2xe^{-y} = 0 \\ (-x^2 + 2y - 4y^2 + y^3)e^{-y} = 0 \end{cases} \iff (x, y) \in \{(0, 0), (0, 2 - \sqrt{2}), (0, 2 + \sqrt{2})\}.$$

On a quatre trois critiques :  $(0, 0)$ ,  $(0, 2 - \sqrt{2})$  et  $(0, 2 + \sqrt{2})$ .

- \* Nature des points critiques :

$$H_f(x, y) = \begin{pmatrix} 2e^{-y} & -2xe^{-y} \\ -2xe^{-y} & (2 + x^2 - 10y + 7y^2 - y^3)e^{-y} \end{pmatrix}, \quad \det(H_f(x, y)) = (4 - 2x^2 - 20y + 14y^2 - 2y^3)e^{-2y}.$$

$\det(H_f(0, 0)) > 0$  et  $\partial_{xx}f(0, 0) > 0$  donc  $(0, 0)$  est un minimum ;

$\det(H_f(0, 2 - \sqrt{2})) < 0$  donc  $(0, 2 - \sqrt{2})$  est un point-selle ;

$\det(H_f(0, 2 + \sqrt{2})) > 0$  et  $\partial_{xx}f(0, 2 + \sqrt{2}) > 0$  donc  $(0, 2 + \sqrt{2})$  est un minimum.

### Exercice 3.14

La société d'Adèle produit deux types d'ampoules : E17 et E24. Indiquons par  $x$  le nombre de milliers d'ampoules de type E17 produites et supposons que la demande pour ce type de lampes est donnée par  $p_1 = 50 - x$ , où  $p_1$  est le prix de vente en euros. De même, indiquons par  $y$  le nombre de milliers d'ampoules de type E24 produites et supposons que la demande pour ce type est donnée par  $p_2 = 60 - 2y$ , où  $p_2$  est aussi le prix de vente en euros. Les coûts communs de production de ces ampoules est  $C = 2xy$  (en milliers d'euros). Par conséquent, le bénéfice de la société d'Adèle (en milliers d'euros) est une fonction de deux variables  $x$  et  $y$ . Déterminer le profit maximal d'Adèle.

#### Correction

La fonction profit en milliers d'euros est  $p(x, y) = p_1x + p_2y - C(x, y) = 50x - x^2 + 60y - 2y^2 - 2xy$ . Pour maximiser le profit, on cherche d'abord les points stationnaires :

$$\nabla p = \mathbf{0} \iff \begin{pmatrix} 50 - 2x - 2y \\ 60 - 4y - 2x \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \begin{cases} x = 20, \\ y = 5. \end{cases}$$

Pour établir la nature de ces points, on étudie la matrice hessienne :

$$\begin{aligned} \partial_{xx}p(x, y) &= -2, & \partial_{xx}p(20, 5) &= -2 < 0, \\ \partial_{xy}p(x, y) &= -2, & \partial_{xy}p(20, 5) &= -2, \\ \partial_{yy}p(x, y) &= -4, & \partial_{yy}p(20, 5) &= -4, \end{aligned}$$

et  $\det(H_f(20, 5)) = (-2)(-4) - (-2)^2 = 4 > 0$  donc  $(20, 5)$  est un point de maximum pour  $p$  et le profit maximal vaut  $p(20, 5) = 650$ . La société d'Adèle réalise le profit maximal de 650000 euros lorsqu'elle vend 20000 ampoules E17 à 30 euros l'une et 5000 ampoules E24 à 50 euros l'une.

### Exercice 3.15

Vous êtes le directeur financier de la firme SANBON & FILS. Cette entreprise a investi 3000 euros pour mettre au point un nouveau parfum. Le coût de la production est de 3 euros par flacon de 100 mL. L'expert consulté par M. SANBON père a établi que si la firme consacre  $x$  euros en publicité pour son parfum et que le prix de vente d'un flacon est de  $y$  euros, la firme vendra exactement  $300 + 6\sqrt{x} - 10y$  pièces. La firme SANBON & FILS fixe évidemment  $x$  et  $y$  de manière à maximiser son profit. En tant que directeur financier, il vous incombe de déterminer ces valeurs.

#### Correction

- \* Revenu de la vente :  $y(300 + 6\sqrt{x} - 10y)$
- \* Coût de production :  $3(300 + 6\sqrt{x} - 10y)$
- \* Coût de développement et de publicité :  $3000 + x$

- \* Profit = (Revenu de la vente) - (Coût de production) - (Coût de développement et de publicité)

Le profit de la firme à maximiser est donc la fonction

$$f: (\mathbb{R}_+^*)^2 \rightarrow \mathbb{R}$$

$$x \mapsto f(x, y) = (y - 3)(300 + 6\sqrt{x} - 10y) - x - 3000$$

La condition nécessaire s'écrit

$$\begin{cases} \partial_x f(x, y) = \frac{3(y-3)}{\sqrt{x}} - 1 = 0 \\ \partial_y f(x, y) = 330 + 6\sqrt{x} - 20y = 0 \end{cases} \implies (x_0, y_0) = (164025, 138).$$

La hessienne en ce point est définie négative :

$$\begin{cases} \partial_{xx} f(x, y) = -\frac{3(y-3)}{2\sqrt{x^3}} \\ \partial_{xy} f(x, y) = \frac{3}{\sqrt{x}} \\ \partial_{yy} f(x, y) = \frac{30(y-3)}{\sqrt{x^3}} - \frac{3}{\sqrt{x}} \end{cases} \implies \det(H_f(x_0, y_0)) = -\frac{241}{32805}.$$

Comme  $\partial_{xx} f(x_0, y_0) = -20$ , on a bien un maximum. La firme SANBON & FILS va donc consacrer 164025 euros à la promotion de son nouveau parfum et vendre le flacon de 100 mL à 138 euros. Elle réalisera de la sorte le profit maximal de  $f(164025, 138) = 15225$  euros.

### Exercice 3.16 (Une fabrication optimale)

Votre société s'occupe de la fabrication d'une pièce mécanique. Celle-ci dépend de deux paramètres réels  $x$  et  $y$  (à priori non-constraints) de la façon suivante : le coût unitaire de fabrication d'une pièce est égal à

$$c(x, y) = x^2 + 2y^2$$

tandis que le taux de pièces défectueuses (compris entre 0 et 1) est égal à

$$t(x, y) = \frac{1}{1 + (xy)^2}.$$

On cherche à maximiser la rentabilité totale du processus de fabrication. On prendra pour fonction objectif le coût unitaire moyen d'une pièce non-défectueuse, qui est égal au coût de fabrication d'une pièce divisé par le taux de pièces non-défectueuses, et on tentera de le simplifier autant que possible.

#### Correction

La fonction à minimiser s'écrit  $f(x, y) = \frac{c(x, y)}{1-t(x, y)} = \frac{x^2 + 2y^2}{1 - \frac{1}{1+(xy)^2}} = \frac{(x^2 + 2y^2)(1+x^2y^2)}{x^2y^2} = \frac{1}{y^2} + x^2 + \frac{2}{x^2} + 2y^2$ . La condition nécessaire s'écrit

$$\begin{cases} \partial_x f(x, y) = 2 \frac{x^4 - 2}{x^3} = 0 \\ \partial_y f(x, y) = 2 \frac{2y^4 - 1}{y^3} = 0 \end{cases} \implies (x_0, y_0) = (\sqrt[4]{2}, 1/\sqrt[4]{2}).$$

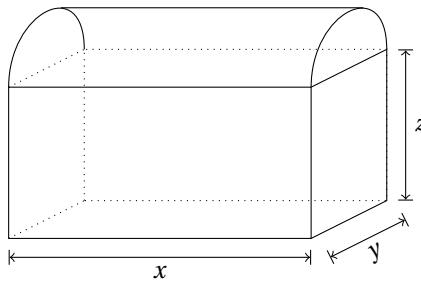
La hessienne en ce point est définie positive :

$$\begin{cases} \partial_{xx} f(x, y) = 2 \frac{x^4 + 6}{x^4} \\ \partial_{xy} f(x, y) = 0 \\ \partial_{yy} f(x, y) = 2 \frac{2y^4 + 3}{y^4} \end{cases} \implies \det(H_f(x_0, y_0)) = 4 \frac{2+6}{2} \frac{1+3}{1/2} > 0.$$

Comme  $\partial_{xx} f(x_0, y_0) > 0$ , on a bien un minimum. En choisissant  $(x, y) = (\sqrt[4]{2}, 1/\sqrt[4]{2})$ , le coût unitaire moyen d'une pièce non-défectueuse est minimale et égal à  $4\sqrt{2}$ .

### Exercice 3.17

Une boîte a la forme d'un parallélépipède surmonté par un demi-cylindre comme dans la figure ci-dessous



On cherche les valeurs  $x, y, z \in \mathbb{R}_+^*$  qui minimisent la surface totale  $S$  de la boîte pour un volume  $V$  égal à  $C$ .

1. Écrire  $S(x, y, z)$
2. Écrire  $V(x, y, z)$
3. Exprimer  $z(x, y)$  comme solution de l'équation  $V(x, y, z) = C$
4. Écrire  $\tilde{S}(x, y) = S(x, y, z(x, y))$ . Calculer et établir la nature des points critiques de  $\tilde{S}(x, y)$

### Correction

1.  $S(x, y, z) = xy + 2xz + 2yz + \pi\left(\frac{y}{2}\right)^2 + \pi\frac{y}{2}x = \left(1 + \frac{\pi}{2}\right)xy + \frac{\pi}{4}y^2 + 2(x + y)z$
2.  $V(x, y, z) = xyz + \frac{1}{2}\pi\left(\frac{y}{2}\right)^2x = xyz + \frac{\pi}{8}xy^2$
3.  $V(x, y, z) = C \iff z = \frac{C - \frac{\pi}{8}xy^2}{xy}$  donc  $z(x, y) = \frac{C}{xy} - \frac{\pi}{8}y$
4.  $\tilde{S}(x, y) = S(x, y, z(x, y)) = \left(1 + \frac{\pi}{2}\right)xy + \frac{\pi}{4}y^2 + 2(x + y)\left(\frac{C}{xy} - \frac{\pi}{8}y\right) = \left(1 + \frac{\pi}{4}\right)xy + \frac{2C}{x} + \frac{2C}{y}$

\* Calcul des points critiques :

$$\nabla \tilde{S}(x, y) = \begin{pmatrix} \left(1 + \frac{\pi}{4}\right)y - \frac{2C}{x^2} \\ \left(1 + \frac{\pi}{4}\right)x - \frac{2C}{y^2} \end{pmatrix} \text{ donc } \nabla \tilde{S}(x, y) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff (x, y) = \left(\sqrt[3]{\frac{2C}{1 + \frac{\pi}{4}}}, \sqrt[3]{\frac{2C}{1 + \frac{\pi}{4}}}\right)$$

Il existe un seul point critique qui est  $\left(\sqrt[3]{\frac{2C}{1 + \frac{\pi}{4}}}, \sqrt[3]{\frac{2C}{1 + \frac{\pi}{4}}}\right)$ .

\* Nature des points critiques :

$$H_{\tilde{S}}(x, y) = \begin{pmatrix} \frac{4C}{x^3} & 1 + \frac{\pi}{4} \\ 1 + \frac{\pi}{4} & \frac{4C}{y^3} \end{pmatrix} \text{ et } \det(H_{\tilde{S}}(x, y)) = \frac{16C^2}{x^3y^3} - \left(1 + \frac{\pi}{4}\right)^2$$

donc

$$H_{\tilde{S}}\left(\sqrt[3]{\frac{2C}{1 + \frac{\pi}{4}}}, \sqrt[3]{\frac{2C}{1 + \frac{\pi}{4}}}\right) = \begin{pmatrix} 2\left(1 + \frac{\pi}{4}\right) & 1 + \frac{\pi}{4} \\ 1 + \frac{\pi}{4} & 2\left(1 + \frac{\pi}{4}\right) \end{pmatrix} \text{ et } \det\left(H_{\tilde{S}}\left(\sqrt[3]{\frac{2C}{1 + \frac{\pi}{4}}}, \sqrt[3]{\frac{2C}{1 + \frac{\pi}{4}}}\right)\right) = 3\left(1 + \frac{\pi}{4}\right)^2.$$

On conclut que l'unique point critique est bien un minimum et l'on a  $z\left(\sqrt[3]{\frac{2C}{1 + \frac{\pi}{4}}}, \sqrt[3]{\frac{2C}{1 + \frac{\pi}{4}}}\right) = \frac{C}{\left(\frac{2C}{1 + \frac{\pi}{4}}\right)^{2/3}} - \frac{\pi}{8}\left(\frac{2C}{1 + \frac{\pi}{4}}\right)^{2/3}$

## Courbe de meilleure approximation

### Exercice 3.18

On considère un ensemble de points expérimentaux  $\{(x_i, y_i)\}_{i=0}^n$  et on suppose que les deux grandeurs  $x$  et  $y$  sont liées, au moins approximativement, par une relation de la forme  $y = a\sin(\frac{\pi}{2}x) + b\cos(\frac{\pi}{2}x)$ . On souhaite alors trouver les constantes  $a$  et  $b$  pour que la courbe d'équation  $y = a\sin(\frac{\pi}{2}x) + b\cos(\frac{\pi}{2}x)$  s'ajuste le mieux possible aux points observés (on parle de *courbe de meilleure approximation*).

Soit  $d_i = y_i - (a\sin(\frac{\pi}{2}x_i) + b\cos(\frac{\pi}{2}x_i))$  l'écart vertical du point  $(x_i, y_i)$  par rapport à la courbe. La méthode de régression (ou des moindres carrés) est celle qui choisit  $a$  et  $b$  de sorte que la somme des carrés de ces déviations soit minimale. Pour cela, on doit minimiser la fonction  $\mathcal{E}$  définie par

$$\begin{aligned} \mathcal{E}: \mathbb{R}^2 &\rightarrow \mathbb{R}_+ \\ (a, b) &\mapsto \mathcal{E}(a, b) = \sum_{i=0}^n d_i^2. \end{aligned}$$

Écrire et résoudre le système linéaire qui permet de calculer  $a$  et  $b$ .

**Correction**

Pour minimiser  $\mathcal{E}$  on cherche ses points stationnaires. Puisque

$$\mathcal{E}(a, b) = \sum_{i=0}^n \left( y_i - \left( a \sin\left(\frac{\pi}{2}x_i\right) + b \cos\left(\frac{\pi}{2}x_i\right) \right) \right)^2$$

calculons tout d'abord les deux dérivées partielles

$$\begin{aligned}\frac{\partial \mathcal{E}}{\partial a}(a, b) &= -2 \left( \sum_{i=0}^n (y_i - (a \sin(\frac{\pi}{2}x_i) + b \cos(\frac{\pi}{2}x_i))) \sin(\frac{\pi}{2}x_i) \right), \\ \frac{\partial \mathcal{E}}{\partial b}(a, b) &= -2 \left( \sum_{i=0}^n (y_i - (a \sin(\frac{\pi}{2}x_i) + b \cos(\frac{\pi}{2}x_i))) \cos(\frac{\pi}{2}x_i) \right),\end{aligned}$$

et cherchons quand elles s'annulent en même temps. On obtient

$$\begin{aligned}\begin{cases} \frac{\partial \mathcal{E}}{\partial a}(a, b) = 0 \\ \frac{\partial \mathcal{E}}{\partial b}(a, b) = 0 \end{cases} &\iff \begin{cases} \sum_{i=0}^n (y_i - (a \sin(\frac{\pi}{2}x_i) + b \cos(\frac{\pi}{2}x_i))) \sin(\frac{\pi}{2}x_i) = 0 \\ \sum_{i=0}^n (y_i - (a \sin(\frac{\pi}{2}x_i) + b \cos(\frac{\pi}{2}x_i))) \cos(\frac{\pi}{2}x_i) = 0 \end{cases} \\ &\iff \begin{cases} \sum_{i=0}^n ((a \sin(\frac{\pi}{2}x_i) + b \cos(\frac{\pi}{2}x_i))) \sin(\frac{\pi}{2}x_i) = \sum_{i=0}^n y_i \sin(\frac{\pi}{2}x_i) \\ \sum_{i=0}^n ((a \sin(\frac{\pi}{2}x_i) + b \cos(\frac{\pi}{2}x_i))) \cos(\frac{\pi}{2}x_i) = \sum_{i=0}^n y_i \cos(\frac{\pi}{2}x_i) \end{cases} \\ &\iff \begin{bmatrix} \sum_{i=0}^n \sin^2(\frac{\pi}{2}x_i) & \sum_{i=0}^n \sin(\frac{\pi}{2}x_i) \cos(\frac{\pi}{2}x_i) \\ \sum_{i=0}^n \sin(\frac{\pi}{2}x_i) \cos(\frac{\pi}{2}x_i) & \sum_{i=0}^n \cos^2(\frac{\pi}{2}x_i) \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^n y_i \sin(\frac{\pi}{2}x_i) \\ \sum_{i=0}^n y_i \cos(\frac{\pi}{2}x_i) \end{bmatrix}.\end{aligned}$$

Si on note

$$U \equiv \sum_{i=0}^n \sin^2(\frac{\pi}{2}x_i), \quad V \equiv \sum_{i=0}^n \sin(\frac{\pi}{2}x_i) \cos(\frac{\pi}{2}x_i), \quad W \equiv \sum_{i=0}^n \cos^2(\frac{\pi}{2}x_i), \quad P \equiv \sum_{i=0}^n y_i \sin(\frac{\pi}{2}x_i), \quad Q \equiv \sum_{i=0}^n y_i \cos(\frac{\pi}{2}x_i),$$

on doit résoudre le système linéaire

$$\begin{pmatrix} U & V \\ V & W \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} P \\ Q \end{pmatrix}$$

dont la solution est

$$a = \frac{WP - VQ}{UW - V^2}, \quad b = \frac{UQ - VP}{UW - V^2}.$$

**Exercice 3.19**

La méthode de régression s'étend facilement à des données qui dépendent de deux ou plusieurs variables. On considère un ensemble de points expérimentaux  $\{(x_i, y_i, z_i)\}_{i=0}^n$  et on suppose que les trois grandeurs  $x$ ,  $y$  et  $z$  sont liées, au moins approximativement, par une relation affine de la forme  $z = a + bx + cy$ . On souhaite alors trouver les constantes  $a$ ,  $b$  et  $c$  pour que le plan d'équation  $z = a + bx + cy$  s'ajuste le mieux possible aux points observés (on parle de *plan de meilleure approximation*).

Soit  $d_i = z_i - (a + bx_i + cy_i)$  l'écart vertical du point  $(x_i, y_i, z_i)$  par rapport au plan. La méthode de régression (ou des moindres carrés) est celle qui choisit  $a$ ,  $b$  et  $c$  de sorte que la somme des carrés de ces déviations soit minimale. Pour cela, on doit minimiser la fonction  $\mathcal{E}$  définie par

$$\begin{aligned}\mathcal{E}: \mathbb{R}^3 &\rightarrow \mathbb{R}_+ \\ (a, b, c) &\mapsto \mathcal{E}(a, b, c) = \sum_{i=0}^n d_i^2.\end{aligned}$$

1. Écrire le système linéaire qui permet de calculer  $a$ ,  $b$  et  $c$
2. Calculer l'équation du plan de meilleure approximation pour l'ensemble  $\{(x_i, y_i, z_i)\}_{i=0}^5$  où

$i$	0	1	2	3	4	5
$x_i$	0	0	1	2	2	2
$y_i$	0	1	0	0	1	2
$z_i$	$\frac{3}{2}$	2	$\frac{1}{2}$	0	$\frac{1}{2}$	1

On utilisera la méthode du pivot de GAUSS pour la résolution du système linéaire.

**Correction**

1. Pour minimiser  $\mathcal{E}$  on cherche ses points stationnaires. Puisque

$$\begin{aligned}\frac{\partial \mathcal{E}}{\partial a}(a, b, c) &= -2 \left( \sum_{i=0}^n (z_i - (a + bx_i + cy_i)) \right), \\ \frac{\partial \mathcal{E}}{\partial b}(a, b, c) &= -2 \left( \sum_{i=0}^n (z_i - (a + bx_i + cy_i)) x_i \right), \\ \frac{\partial \mathcal{E}}{\partial c}(a, b, c) &= -2 \left( \sum_{i=0}^n (z_i - (a + bx_i + cy_i)) y_i \right),\end{aligned}$$

on obtient

$$\begin{aligned}\begin{cases} \frac{\partial \mathcal{E}}{\partial a}(a, b, c) = 0 \\ \frac{\partial \mathcal{E}}{\partial b}(a, b, c) = 0 \\ \frac{\partial \mathcal{E}}{\partial c}(a, b, c) = 0 \end{cases} &\iff \begin{cases} \sum_{i=0}^n (z_i - (a + bx_i + cy_i)) = 0 \\ \sum_{i=0}^n (z_i - (a + bx_i + cy_i)) x_i = 0 \\ \sum_{i=0}^n (z_i - (a + bx_i + cy_i)) y_i = 0 \end{cases} \iff \begin{cases} \sum_{i=0}^n (a + bx_i + cy_i) = \sum_{i=0}^n z_i \\ \sum_{i=0}^n (ax_i + bx_i^2 + cy_i x_i) = \sum_{i=0}^n z_i x_i \\ \sum_{i=0}^n (ay_i + bx_i y_i + cy_i^2) = \sum_{i=0}^n z_i y_i \end{cases} \\ &\iff \begin{pmatrix} (n+1) & \sum_{i=0}^n x_i & \sum_{i=0}^n y_i \\ \sum_{i=0}^n x_i & \sum_{i=0}^n x_i^2 & \sum_{i=0}^n x_i y_i \\ \sum_{i=0}^n y_i & \sum_{i=0}^n x_i y_i & \sum_{i=0}^n y_i^2 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^n z_i \\ \sum_{i=0}^n z_i x_i \\ \sum_{i=0}^n z_i y_i \end{pmatrix}.\end{aligned}$$

2. Dans notre cas,

$$\begin{array}{lcl} \sum_{i=0}^n x_i = 7 & \sum_{i=0}^n y_i = 4 & \sum_{i=0}^n z_i = \frac{11}{2} \\ \sum_{i=0}^n x_i y_i = 6 & \sum_{i=0}^n x_i z_i = \frac{7}{2} & \sum_{i=0}^n y_i z_i = \frac{9}{2} \\ n+1 = 6 & \sum_{i=0}^n x_i^2 = 13 & \sum_{i=0}^n y_i^2 = 6 \end{array}$$

donc on a le système linéaire

$$\begin{pmatrix} 6 & 7 & 4 \\ 7 & 13 & 6 \\ 4 & 6 & 6 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 11/2 \\ 7/2 \\ 9/2 \end{pmatrix}$$

qu'on peut résoudre par la méthode de GAUSS

$$\left( \begin{array}{ccc|c} 6 & 7 & 4 & 11/2 \\ 7 & 13 & 6 & 7/2 \\ 4 & 6 & 6 & 9/2 \end{array} \right) \xrightarrow{L_2 \leftarrow L_2 - \frac{7}{6}L_1} \left( \begin{array}{ccc|c} 6 & 7 & 4 & 11/2 \\ 0 & 29/6 & 4/3 & -35/12 \\ 0 & 4/3 & 10/3 & 5/6 \end{array} \right) \xrightarrow{L_3 \leftarrow L_3 - \frac{8}{29}L_2} \left( \begin{array}{ccc|c} 6 & 7 & 4 & 11/2 \\ 0 & 29/6 & 4/3 & -35/12 \\ 0 & 0 & 86/29 & 95/58 \end{array} \right)$$

dont la solution est

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 123/86 \\ -65/86 \\ 95/172 \end{pmatrix} \approx \begin{pmatrix} 1.430232557 \\ -0.7558139503 \\ 0.5523255766 \end{pmatrix}.$$

**★ Exercice 3.20**

On se propose d'écrire une `function` pour évaluer le polynôme de fitting d'un ensemble de points. Chaque `function` prend en entrée `P` une matrice de  $n$  lignes et 2 colonnes qui contient les points d'interpolation, `x` le vecteur contenant les points où on veut évaluer le polynôme de fitting et `m` le degré du polynôme de fitting et elle donne en sortie `y` le vecteur contenant l'évaluation du polynôme de fitting.

Compléter la `function` suivante

```
function [y]=fittingpolynomial(P,x,m)
[1,c]=size(P);
for r=1:m+1
    V(r,1) =
    b(r)=
end
for c=2:m+1
    for r=1:m
```

```

    V(r,c) = V(r+1,c-1);
end
V(m+1,c) =
end
alpha = V\b';
y=zeros(size(x));
for i=1:m+1
    y+=alpha(i)*x.^^(i-1);
end
end

```

et la tester en comparant le fitting linéaire, le fitting parabolique et l'interpolation sur un jeu de  $n = 3$  points (donc pour  $m = 2$  le fitting retrouve le polynôme d'interpolation) ; puis comparer le fitting linéaire, le fitting parabolique et l'interpolation sur le jeu de points suivant :

```

P=[1 6.008; 2.5 15.722; 3.5 27.130 ; 4 33.772; 1.1 5.257; 1.8 9.549; 2.2 11.098];
x=[1:0.1:4];
ylin=fittingpolynomial(P,x,1);
ypar=fittingpolynomial(P,x,2);
ynew=newton(P,x)
plot(P(:,1),P(:,2),'.o',x,ylin,x,ypar,x,ynew,'. ');

```

## Correction

Dans le fichier `fittingpolynomial.m` on définit la fonction suivante

```

function [y]=fittingpolynomial(P,x,m)
[l,c]=size(P);
for r=1:m+1
    V(r,1) = sum( P(:,1).^(r-1) );
    b(r)=sum( P(:,2).*P(:,1)).^(r-1) );
end
for c=2:m+1
    for r=1:m
        V(r,c) = V(r+1,c-1) ;
    end
    V(m+1,c) = sum( P(:,1).^(m+c-1) ) ;
end
alpha = V\b';
y=zeros(size(x));
for i=1:m+1
    y+=alpha(i)*x.^^(i-1);
end
end

```

On peut décomposer notre fonction en deux fonctions : la première rend les coefficient du polynôme dans la base canonique, la deuxième évalue le polynôme lorsqu'on connaît ces coefficients :

```

function [alpha]=fittingpolynomialPoly(P,m)
[l,c]=size(P);
for r=1:m+1
    V(r,1) = sum( P(:,1).^(r-1) );
    b(r)=sum( P(:,2).*P(:,1)).^(r-1) );
end
for c=2:m+1
    for r=1:m
        V(r,c) = V(r+1,c-1) ;
    end
    V(m+1,c) = sum( P(:,1).^(m+c-1) ) ;
end
alpha = V\b';
end

```

```

function [y]=fittingpolynomialEval(alpha,x,m)
y=zeros(size(x));
for i=1:m+1
    y+=alpha(i)*x.^^(i-1);
end

```

On écrit le script de test

```
% Points d'interpolation : P(i)=(x(i),y(i))
```

```

P=[-2 4; 0 0; 1 1];

% Pour l'affichage on evaluera les polynomes en les points suivants
x=[-2:.1:2];

% le seul polynome de degre 2 qui interpole ces points est la parabole d'equation y=x^2
ynew=newton(P,x);

% polynome de degre 1 qui fitte ces points
alphalin=fittingpolynomialPoly(P,1)
ylin=fittingpolynomialEval(alphalin,x,1);
% polynome de degre 2 qui fitte ces points
alphapar=fittingpolynomialPoly(P,2)
ypar=fittingpolynomialEval(alphalin,x,2);

figure(1)
plot(P(:,1),P(:,2), 'o',x,ylin,x,ypar,x,ynew, ':');

% comparons avec les polynomes calcules directement par Octave
figure(2)
olin=polyval(polyfit(P(:,1),P(:,2),1),x);
opar=polyval(polyfit(P(:,1),P(:,2),2),x);
plot(P(:,1),P(:,2), 'o',x,olin,x,opar);

```

puis le script de test

```

P=[1 6.008; 2.5 15.722; 3.5 27.130 ; 4 33.772; 1.1 5.257; 1.8 9.549; 2.2 11.098];
x=[1:0.1:4];

% n=7; polynome de degre n-1=6 interpolant ces points
ynew=newton(P,x)

% polynome de degre 1 qui fitte ces points
alphalin=fittingpolynomialPoly(P,1)
ylin=fittingpolynomialEval(alphalin,x,1);
% polynome de degre 2 qui fitte ces points
alphapar=fittingpolynomialPoly(P,2)
ypar=fittingpolynomialEval(alphalin,x,2);

plot(P(:,1),P(:,2), 'o',x,ylin,x,ypar,x,ynew, ':');

```

### ★ Exercice 3.21 (Interpolation et fitting, utilisation de fonction vues en TP)

Dans cet exercice on fera appel à la fonction `fittingpolynomial.m` écrite en TP pour calculer le fitting polynomial et à l'une des fonction écrites en TP pour calculer le polynôme d'interpolation, à savoir `naive.m` ou `lagrange.m` ou `newton.m`.

- a) Dans le fichier `exercice2a.m` écrire un **script** qui compare sur un même graphique le fitting linéaire, le fitting parabolique et le polynôme d'interpolation sur le jeu de points suivant :

$$\{(-1, 1), (0, 0), (1, 1), (2, 8)\}.$$

Afficher à l'écran les coefficients de ces polynômes.

- b) Dans le fichier `exercice2b.m` écrire un **script** qui répète le même exercice pour le jeu de points suivant

$$\{(-1, 1), (0, 0), (1, 1), (2, 4)\}$$

et commenter ce deuxième résultat.

Conseil : vérifiez vos programmes en donnant plusieurs exemples pertinents d'utilisation de vos fonctions et en vous donnant les moyens de le vérifier. Comparez lorsque c'est possible votre programme aux réponses des fonctions d'Octave. Par exemple, la commande `polyval(polyfit(P(:,1),P(:,2),deg),x);` doit renvoyer le même résultat que `fittingpolynomial(P,x,deg)`.

#### Correction

En TP, dans les fichiers `fittingpolynomialPoly.m` et `naivePoly.m` on a écrit les fonction

```
function [alpha]=fittingpolynomialPoly(P,m)
V(1:m+1,1) = sum( P(:,1).^(0:m) );
b(1:m+1)=sum( P(:,2).*P(:,1)).^(0:m) );
for c=2:m+1
    V(1:m,c) = V(2:m+1,c-1);
    V(m+1,c) = sum( P(:,1).^(m+c-1) );
end
alpha = V\b';
end
```

```
function [alpha]=naivePoly(P)
[l,c]=size(P);
V = ones(l,1);
V(:,2:l) = P(:,1).^(1:l-1);
alpha = V\P(:,2);
end
```

qui calculent les coefficients des polynômes de régression et d'interpolation.

Ensuite, dans les fichiers `fittingpolynomialEval.m` et `naiveEval.m` on a écrit les fonction pour évaluer les polynômes ainsi construit en un vecteur de points :

```
function [y]=fittingpolynomialEval(alpha,x)
y=zeros(size(x));
for i=1:length(alpha)
    y+=alpha(i)*x.^^(i-1);
end
end
```

```
function [y]=naiveEval(alpha,x)
y=zeros(size(x));
for k=size(alpha):-1:1
    y=alpha(k)+x.*y;
end
end
```

a) Dans le fichier `exercice2a.m` on écrit le script

```
P=[-1 1; 0 0; 1 1; 2 8];

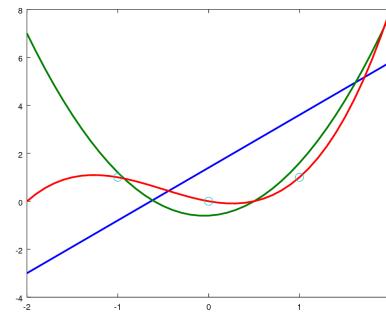
alphaLin=fittingpolynomialPoly(P,1)
alphaPar=fittingpolynomialPoly(P,2)
alphaInt=naivePoly(P)

x=[-2:0.1:2];

ylin=fittingpolynomialEval(alphaLin,x);
ypar=fittingpolynomialEval(alphaPar,x);
yinterpol=naiveEval(alphaInt,x);

plot(x,ylin,'LineWidth',2,...
      x,ypar,'LineWidth',2,...
      x,yinterpol,'LineWidth',2,...)
```

```
P(:,1),P(:,2),'o','MarkerSize',10);
saveas (gcf, "2016-cc1-exo2a.png", "png");
```



En effet,

$$\begin{aligned} p_{\text{linéaire}}(x) &= \alpha_0 + \alpha_1 x \\ p_{\text{parabolique}}(x) &= \beta_0 + \beta_1 x + \beta_2 x^2 \\ p_{\text{interpolation}}(x) &= \frac{x(x-1)(x-2)}{-6} + \frac{(x+1)x(x-2)}{-2} + 8 \frac{(x+1)x(x-1)}{6} \end{aligned}$$

avec

$$\begin{pmatrix} 4 & 2 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} = \begin{pmatrix} 10 \\ 16 \end{pmatrix} \quad \begin{pmatrix} 4 & 2 & 6 \\ 2 & 6 & 8 \\ 6 & 8 & 18 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 10 \\ 16 \\ 34 \end{pmatrix}$$

ainsi  $\alpha_0 = \frac{7}{5}$ ,  $\alpha_1 = \frac{6}{5}$ ,  $\beta_0 = -\frac{3}{5}$ ,  $\beta_1 = \frac{1}{5}$ ,  $\beta_2 = 2$ .

b) Dans le fichier `exercice2b.m` on écrit le script

```
P=[-1 1; 0 0; 1 1; 2 4];

alphaLin=fittingpolynomialPoly(P,1)
alphaPar=fittingpolynomialPoly(P,2)
alphaInt=naivePoly(P)
```

```
x=[-2:0.1:2];

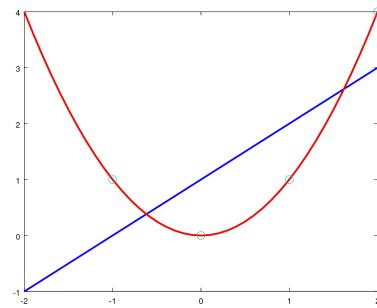
ylin=fittingpolynomialEval(alphaLin,x);
ypar=fittingpolynomialEval(alphaPar,x);
```

```

yinterpol=naiveEval(alphaInt,x);

plot(x,ylin,'LineWidth',2, ...
      x,ypar,'LineWidth',2, ...
      x,yinterpol,'LineWidth',2, ...
      P(:,1),P(:,2),'o','MarkerSize',10);
saveas(gcf, "2016-cc1-exo2b.png", "png");

```



En effet,

$$\begin{aligned}
 p_{\text{linéaire}}(x) &= \alpha_0 + \alpha_1 x \\
 p_{\text{parabolique}}(x) &= \beta_0 + \beta_1 x + \beta_2 x^2 \\
 p_{\text{interpolation}}(x) &= \frac{x(x-1)(x-2)}{-6} + \frac{(x+1)x(x-2)}{-2} + 4 \frac{(x+1)x(x-1)}{6} = x^2
 \end{aligned}$$

avec

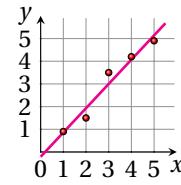
$$\begin{pmatrix} 4 & 2 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} = \begin{pmatrix} 6 \\ 8 \end{pmatrix} \quad \begin{pmatrix} 4 & 2 & 6 \\ 2 & 6 & 8 \\ 6 & 8 & 18 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 6 \\ 8 \\ 18 \end{pmatrix}$$

ainsi  $\alpha_0 = 1$ ,  $\alpha_1 = 1$ ,  $\beta_0 = 0$ ,  $\beta_1 = 0$ ,  $\beta_2 = 1$ . Dans ce cas, le fitting parabolique coïncide avec le polynôme d'interpolation car ce dernier n'appartient pas simplement à  $\mathbb{R}_3[x]$  mais à  $\mathbb{R}_2[x]$ .

### Exercice 3.22

Calculer la droite de meilleure approximation de l'ensemble de points suivant :

$x$	1	2	3	4	5
$y$	0.9	1.5	3.5	4.2	4.9



### Correction

il s'agit de chercher  $a_0$  et  $a_1$  qui minimisent l'erreur  $\mathcal{E}(a_0, a_1) = \sum_{i=0}^2 (y_i - (a_0 + a_1 x_i))^2$ . Cela impose la résolution du système linéaire

$$\begin{pmatrix} (n+1) & \sum_{i=0}^n x_i \\ \sum_{i=0}^n x_i & \sum_{i=0}^n x_i^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^n y_i \\ \sum_{i=0}^n y_i x_i \end{pmatrix} \implies \begin{pmatrix} 6 & 15 \\ 15 & 55 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 15 \\ 55.7 \end{pmatrix}$$

Donc  $a_0 = -0.21$  et  $a_1 = 1.07$ .

En utilisant les fonction de l'exercice 3.20 on écrit le script de test

```

% Points d'interpolation : P(i)=(x(i),y(i))
Px=[1:5];
Py=[ 0.9; 1.5; 3.5; 4.2; 4.9];
P=[Px Py]

% Pour l'affichage on evaluera les polynomes en les points suivants
x=[0:.1:6];

% polynome de degre 1 qui fitte ces points
ylin=fittingpolynomial(P,x,1);
% polynome de degre 2 qui fitte ces points
ypar=fittingpolynomial(P,x,2);

figure(1)
plot(P(:,1),P(:,2),'o',x,ylin,x,ypar);

% comparons avec les polynomes calcules directement par Octave

```

```
figure(2)
olin=polyval(polyfit(P(:,1),P(:,2),1),x);
opar=polyval(polyfit(P(:,1),P(:,2),2),x);
plot(P(:,1),P(:,2),'o',x,olin,x,opar);
```

### Exercice 3.23

À partir des données

$x$	1.0	2.5	3.5	4.0	1.1	1.8	2.2	3.7
$y$	6.008	15.722	27.130	33.772	5.257	9.549	11.098	28.828

on veut calculer la droite et la parabole de régression et comparer les erreurs des chaque régression.

#### Correction

- La droite de régression a équation  $y = mx + q$  avec

$$m = \frac{\sum_{i=0}^6 y_i(x_i - \bar{x})}{\sum_{i=0}^7 x_i(x_i - \bar{x})} \approx 8.420042377, \quad q = \bar{y} - m\bar{x} \approx -3.37833528,$$

où  $\bar{x} = \frac{1}{7} \sum_{i=0}^6 x_i = 2.0125$  et  $\bar{y} = \frac{1}{7} \sum_{i=0}^6 y_i = 13.567$ . L'erreur est

$$\sum_{i=0}^6 (y_i - (mx_i + q))^2 = 39.59960820.$$

- La parabole de régression a équation  $y = a_0 + a_1x + a_2x^2$  avec  $a_0, a_1, a_2$  solution du système linéaire

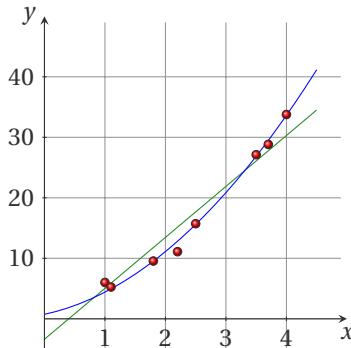
$$\begin{pmatrix} 8 & \sum_{i=0}^6 x_i & \sum_{i=0}^6 x_i^2 \\ \sum_{i=0}^6 x_i & \sum_{i=0}^6 x_i^2 & \sum_{i=0}^6 x_i^3 \\ \sum_{i=0}^6 x_i^2 & \sum_{i=0}^6 x_i^3 & \sum_{i=0}^6 x_i^4 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^6 y_i \\ \sum_{i=0}^6 y_i x_i \\ \sum_{i=0}^6 y_i x_i^2 \end{pmatrix} \quad \text{i.e.} \quad \begin{pmatrix} 8 & 16.1 & 44.79 \\ 16.1 & 44.79 & 141.311 \\ 44.79 & 141.311 & 481.5123 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 108.536 \\ 322.7425 \\ 1067.97905 \end{pmatrix}$$

et on obtient

$$\begin{cases} a_0 = 0.744611871628180655, \\ a_1 = 2.14480468957977077, \\ a_2 = 1.51926210146774388. \end{cases}$$

L'erreur est

$$\sum_{i=0}^6 (y_i - (a_0 + a_1 x_i + a_2 x_i^2))^2 = 5.715921703.$$



### Exercice 3.24

Le tableau ci-dessous donne la conductivité thermique  $k$  du sodium pour différentes valeurs de la température. On veut calculer la parabole de meilleure approximation.

$T$ (°C)	79	190	357	524	690
$k$	1.00	0.932	0.839	0.759	0.693

**Correction**

La parabole de régression a équation  $y = a_0 + a_1 x + a_2 x^2$  avec  $a_0, a_1, a_2$  solution du système linéaire

$$\begin{pmatrix} 6 & \sum_{i=0}^4 x_i & \sum_{i=0}^4 x_i^2 \\ \sum_{i=0}^4 x_i & \sum_{i=0}^4 x_i^2 & \sum_{i=0}^4 x_i^3 \\ \sum_{i=0}^4 x_i^2 & \sum_{i=0}^4 x_i^3 & \sum_{i=0}^4 x_i^4 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^4 y_i \\ \sum_{i=0}^4 y_i x_i \\ \sum_{i=0}^4 y_i x_i^2 \end{pmatrix} \quad i.e. \quad \begin{pmatrix} 8 & 16.1 & 44.79 \\ 16.1 & 44.79 & 141.311 \\ 44.79 & 141.311 & 481.5123 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 108.536 \\ 322.7425 \\ 1067.97905 \end{pmatrix}$$

et on obtient

$$\begin{cases} a_0 = 0.744611871628180655, \\ a_1 = 2.14480468957977077, \\ a_2 = 1.51926210146774388. \end{cases}$$

L'erreur est

$$\sum_{i=0}^6 (y_i - (a_0 + a_1 x_i + a_2 x_i^2))^2 = 5.715921703.$$

**Exercice 3.25**

La viscosité cinématique  $\mu$  de l'eau varie en fonction de la température comme dans le tableau suivant :

$T$ (°C)	0	21.1	37.8	54.4	71.1	87.8	100
$\mu$ ( $10^{-3} \text{ m}^2 \cdot \text{s}^{-1}$ )	1.79	1.13	0.696	0.519	0.338	0.321	0.296

On veut évaluer les valeurs  $\mu(10^\circ)$ ,  $\mu(30^\circ)$ ,  $\mu(60^\circ)$ ,  $\mu(90^\circ)$  par le polynôme de meilleur approximation de degré 3.

**Correction**

On a la famille de points  $\{(T_i, \mu_i)\}_{i=0}^6$ . Le polynôme de meilleur approximation de degré 3 s'écrit

$$r(T) = a_0 + a_1 T + a_2 T^2 + a_3 T^3$$

où  $a_0, a_1, a_2, a_3$  sont solution du système linéaire

$$\begin{pmatrix} 6 & \sum_{i=0}^6 T_i & \sum_{i=0}^6 T_i^2 & \sum_{i=0}^6 T_i^3 \\ \sum_{i=0}^6 T_i & \sum_{i=0}^6 T_i^2 & \sum_{i=0}^6 T_i^3 & \sum_{i=0}^6 T_i^4 \\ \sum_{i=0}^6 T_i^2 & \sum_{i=0}^6 T_i^3 & \sum_{i=0}^6 T_i^4 & \sum_{i=0}^6 T_i^5 \\ \sum_{i=0}^6 T_i^3 & \sum_{i=0}^6 T_i^4 & \sum_{i=0}^6 T_i^5 & \sum_{i=0}^6 T_i^6 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^6 \mu_i \\ \sum_{i=0}^6 \mu_i T_i \\ \sum_{i=0}^6 \mu_i T_i^2 \\ \sum_{i=0}^6 \mu_i T_i^3 \end{pmatrix}$$

et on obtient

$$\begin{cases} a_0 = 0.914534618675843625, \\ a_1 = 0.914534618675843625, \\ a_2 = -0.000620138768106035594, \\ a_3 = -0.000620138768106035594. \end{cases}$$

On a alors

$$r(10^\circ) = 1.004300740 \quad r(30^\circ) = 0.9114735501 \quad r(60^\circ) = 0.9114735501 \quad r(90^\circ) = 0.249145396$$

**Exercice 3.26**

Considérons  $n = 10$  points  $P_i = (x_i, y_i)$  avec  $x_i = (i - 1)/(n - 1)$ ,  $i = 1, \dots, n$  et  $y_i = 2x_i + 1 + \epsilon_i$  avec  $\epsilon_i \in ]0; 0.01[$  généré aléatoirement avec une distribution normale. Comparer l'interpolation et le fitting linéaire sur ce jeu de points.

**Correction**

On écrira le script suivant (on compare les résultats de nos `function` avec ceux issues des `function` déjà implémentées dans Octave) :

```
% Points d'interpolation : P(i)=(x(i),y(i))
n = 10
P=zeros(n,2);
P(:,1)=linspace(0,2,n);
P(:,2)=2*P(:,1)+1+0.1*randn(n,1);

% Pour l'affichage on évaluera les polynômes en les points suivants
```

```

x=[0:.01:2];

% polynome de degre n-1 interpolant ces points
ynew=newton2(P,x);
% le meme calcule directement par Octave
ointerp=polyval(polyfit(P(:,1),P(:,2),n-1),x); %=ynew

% polynome de degre 1 qui fitte ces points
ylin=fittingpolynomial(P,x,1);
% le meme calcule directement par Octave
olin=polyval(polyfit(P(:,1),P(:,2),1),x); %=ylin

%
subplot(1,2,1)
plot(P(:,1),P(:,2), 'o', x,ylin, 'r-', 'LineWidth', 2, x,ynew, 'b:', 'LineWidth', 2);
title("Nos fonctions")
axis ([0, 2, min(P(:,2)), max(P(:,2))])

subplot(1,2,2)
plot(P(:,1),P(:,2), 'o', x,olin, 'r-', 'LineWidth', 2,x,ointerp, 'b:', 'LineWidth', 2);
title("Calcul Octave")
axis ([0, 2, min(P(:,2)), max(P(:,2))])

```

### ★ Exercice 3.27

L'espérance de vie dans un pays a évolué dans le temps selon le tableau suivant :

	Année	1975	1980	1985	1990
Espérance	72,8	74,2	75,2	76,4	

Utiliser l'interpolation polynomiale pour estimer l'espérance de vie en 1977, 1983 et 1988. La comparer avec une interpolation linéaire par morceaux et avec un fitting polynomiale avec  $m = 1, 2$  (avec  $m = 3$  on retrouve le polynôme d'interpolation).

### Correction

Si on choisit de poser  $x_0 = 0$  pour l'année 1975,  $x_1 = 5$  pour l'année 1980 etc., on construit

- ★  $p_1 \in \mathbb{R}_1[x]$  la droite de meilleure approximation (fitting  $m = 1$ )
- ★  $p_2 \in \mathbb{R}_2[x]$  la parabole de meilleure approximation (fitting  $m = 2$ )
- ★  $p_3 \in \mathbb{R}_3[x]$  le polynôme d'interpolation ( $n = 3$ )
- ★  $s_1$  la spline linéaire

et on évalue ces fonctions en  $x = 2, 8$  et  $13$  (notons que seuls  $p_3$  et  $s_1$  interpolent les données) :

```

Px=[0:5:15];
Py=[72.8 74.2 75.2 76.4];

n=length(Px)-1;

p1=polyfit(Px,Py,1);
p2=polyfit(Px,Py,2);
p3=polyfit(Px,Py,n); % fitting avec m=n equivalent a interpolation

x=[2 8 13];
y1=polyval(p1,x)
y2=polyval(p2,x)
y3=polyval(p3,x)
for j=1:3
    s1(j)=polyval(polyfit(Px(j:j+1),Py(j:j+1),1),x(j));
end
s1

% Plots
xx=[0:.1:15];
y1=polyval(p1,xx);
y2=polyval(p2,xx);

```

```

y3=polyval(p3,xx);
% for j=1:3
% s1(j)=polyval(polyfit(Px(j,j+1),Py(j,j+1),1),xx(j))
% end

plot(Px,Py,'DisplayName','Points','o','MarkerFaceColor',[.49 1 .63],'MarkerSize',10, ...
    xx,y1,'DisplayName','p_1','r','LineWidth',2, ...
    xx,y2,'DisplayName','p_2','b','LineWidth',2, ...
    xx,y3,'DisplayName','p_3','m','LineWidth',2, ...
    Px,Py,'DisplayName','s_1','k','LineWidth',2);
legend('show','Location','northwest','boxoff')
set(gca,'XTick',[0,2,5,8,10,13,15]);
axis ([0 15 72 77]);
grid;

```

On obtient les estimations suivantes :

	Année	1977	1983	1988
Espérance $p_1$	73.352	74.768	75.948	
Espérance $p_2$	73.354	74.830	75.950	
Espérance $p_3$	73.446	74.810	75.858	
Espérance $s_1$	73.360	74.800	75.920	

### ★ Exercice 3.28

Considérons l'ensemble de points  $\{(x_i, y_i)\}_{i=0}^n$  ainsi construit :  $x_i \in [-1; 1]$  tous distincts et  $y_i = (x_i - 1)x_i(x_i + 1) + r_i$  où  $r_i \in [-0.6; 0.6]$  peut être considéré comme un bruit aléatoire associé au signal  $(x_i - 1)x_i(x_i + 1)$ . On cherche une fonction d'équation  $y = \sum_{j=0}^m a_j \phi_j(x_i)$  qui approche au mieux cet ensemble de points. Pour cela, on doit minimiser la fonction  $\mathcal{E}: \mathbb{R}^{m+1} \rightarrow \mathbb{R}_+$  définie par

$$\mathcal{E}(a_0, a_1, \dots, a_m) = \sum_{i=0}^n \left( y_i - \sum_{j=0}^m a_j \phi_j(x_i) \right)^2.$$

Il faut alors fixer  $m$  et résoudre le système linéaire  $\mathbb{A}^T \mathbb{A} \mathbf{a} = \mathbb{A}^T \mathbf{y}$  de  $(m + 1)$  équations en les  $(m + 1)$  inconnues  $a_0, a_1, \dots, a_m$  avec  $\mathbf{y} = (y_0, y_1, \dots, y_n)$  et

$$\mathbb{A} \stackrel{\text{def}}{=} \underbrace{\begin{pmatrix} \phi_0(x_0) & \phi_1(x_0) & \dots & \phi_m(x_0) \\ \phi_0(x_1) & \phi_1(x_1) & \dots & \phi_m(x_1) \\ \vdots & \vdots & & \vdots \\ \phi_0(x_n) & \phi_1(x_n) & \dots & \phi_m(x_n) \end{pmatrix}}_{n \times m}.$$

Choisissons les fonctions  $\phi_j$  de la forme  $\phi_j(x) = \sin(j\pi x)$ ,  $j = 0, \dots, m$ . Soit  $n = 20$  et  $m = 2$ . Comparer sur un graphe la fonction  $y = (x - 1)x(x + 1)$  (le signal souhaité), les points (le signal bruité) et la fonction de meilleure approximation (le signal lissé). Estimer l'erreur entre le signal souhaité et le signal lissé. Répéter le même exercice pour  $n = 20$  et  $m = 8$ , puis  $n = 200$  et  $n = 8$ .

### Correction

```

clear all
% signal
f=@(x) [(x-1).*(x+1).*x];
% signal bruite
n=20;
P=zeros(n,3);
P(:,1)=sort(2*rand(n,1)-1); %linspace(-1,1,n);
%P(1,1)=-1;
%P(n,1)=1;
P(:,2)=f(P(:,1)); %sin(6*pi*P(:,1)); % signal
P(:,3)=P(:,2)+(2.*rand(n,1)-1)*0.6; % ajout du bruit

% calcul du signal lisse
m=4;
phi=@(k,xi)[sin(k*pi*xi)];
for j=1:m

```

```

A(:,j)=phi(j,P(:,1));
end
alpha = (A'*A)\(A'*P(:,2));

% AFFICHAGE
x=[-1:0.1:1];

% evaluation du signal lisse
ylisse=zeros(size(x));
for i=1:m
    ylisse.+=alpha(i)*phi(i,x);
end

% evaluation du signal initial
yinit=f(x);

xmin=min(x);
xmax=max(x);
ymin=min(P(:,3));
ymax=max(P(:,3));

subplot(2,2,1)
plot(x,yinit)
axis([xmin xmax ymin ymax]);
title('Signal initial')

subplot(2,2,2)
plot(P(:,1),P(:,3),'o')
axis([xmin xmax ymin ymax]);
title('Signal bruite')

subplot(2,2,3)
plot(P(:,1),P(:,3),'o',x,ylisse,'r- ',x,yinit,'g--')
axis([xmin xmax ymin ymax]);
title('Signal lisse, bruite et initial')

subplot(2,2,4)
plot(x,abs(yinit-ylisse))
title('|initial-lisse|')

```



# Chapitre 4

## De l'interpolation à l'approximation d'EDO

### 4.1 Méthodes de quadrature interpolatoires

On sait bien qu'il n'est pas toujours possible, pour une fonction arbitraire, de trouver la forme explicite d'une primitive. Par exemple, comment peut-on tracer le graphe de la fonction erf (appelée fonction d'erreur de GAUSS) définie comme suit ?

$$\text{erf}: \mathbb{R} \rightarrow \mathbb{R}$$

$$x \mapsto \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

Mais même quand on la connaît, il est parfois difficile de l'utiliser. C'est par exemple le cas de la fonction  $f(x) = \cos(4x) \cos(3 \sin(x))$  pour laquelle on a

$$\int_0^\pi f(x) dx = \pi \sum_{k=0}^{\infty} \frac{(-9/4)^k}{k!(k+4)!};$$

on voit que le calcul de l'intégrale est transformé en un calcul, aussi difficile, de la somme d'une série. Dans certains cas, la fonction à intégrer n'est connue que par les valeurs qu'elle prend sur un ensemble fini de points (par exemple, des mesures expérimentales). On se trouve alors dans la même situation que celle abordée au chapitre précédent pour l'approximation des fonctions. Dans tous ces cas, il faut considérer des méthodes numériques afin d'approcher la quantité à laquelle on s'intéresse, indépendamment de la difficulté à intégrer la fonction.

Dans les méthodes d'intégration, l'intégrale d'une fonction  $f$  continue sur un intervalle borné  $[a, b]$  est remplacée par une somme finie. Le choix des nœuds et celui des coefficients qui interviennent dans la somme approchant l'intégrale sont des critères essentiels pour minimiser l'erreur.

#### 4.1.1 Principes généraux

Soit  $f$  une fonction réelle intégrable sur l'intervalle  $[a; b]$ . Le calcul explicite de l'intégrale définie  $I_{[a;b]}(f) \equiv \int_a^b f(x) dx$  peut être difficile, voire impossible. On appelle *formule de quadrature* ou *formule d'intégration numérique* toute formule permettant de calculer une approximation de  $I_{[a;b]}(f)$ . Une possibilité consiste à remplacer  $f$  par une approximation  $\tilde{f}$  et calculer  $I_{[a;b]}(\tilde{f})$  au lieu de  $I_{[a;b]}(f)$ . L'approximation  $\tilde{f}$  doit être facilement intégrable, ce qui est le cas si, par exemple,  $\tilde{f}$  est un polynôme. Typiquement le calcul explicite de l'intégrale de  $\tilde{f}$  sur  $[a; b]$  permet d'écrire la formule de quadrature sous la forme

$$I_{[a;b]}(\tilde{f}) = \sum_{i=0}^n \alpha_i \tilde{f}(x_i)$$

qui est une somme pondérée des valeurs de  $\tilde{f}$  sur un ensemble de  $n + 1$  nœuds distincts  $\{x_i\}_{i=0}^{i=n}$  : on dit que ces points sont les *nœuds* de la formule de quadrature et que les nombres  $\alpha_i \in \mathbb{R}$  sont les *coefficients* ou encore les *poids*.

Si  $f$  est de classe  $\mathcal{C}^0$  sur  $[a; b]$ , l'erreur de quadrature  $E_{[a;b]}(f) \equiv |I_{[a;b]}(f) - I_{[a;b]}(\tilde{f})|$  satisfait

$$E_{[a;b]}(f) = \left| \int_a^b f(x) - \tilde{f}(x) dx \right| \leq \int_a^b |f(x) - \tilde{f}(x)| dx \leq (b - a) \max_{x \in [a;b]} |f(x) - \tilde{f}(x)|.$$

#### 4.1.2 Exemples de formules de quadrature interpolatoires

Une approche naturelle pour construire une formule de quadrature consiste à prendre pour  $p_n$  le polynôme d'interpolation de  $f$  sur un ensemble de  $n + 1$  noeuds distincts  $\{x_i\}_{i=1}^{i=n+1}$ .

Nom	Points d'interpolation	Polynôme $p_n \in \mathbb{R}_n[t]$	$\int_a^b p_n(t) dt$
Rectangle à gauche	$a$	$p_0(t) = f(a) \in \mathbb{R}_0[t]$	$(b-a)f(a)$
Rectangle à droite	$b$	$p_0(t) = f(b) \in \mathbb{R}_0[t]$	$(b-a)f(b)$
Point milieu	$\frac{a+b}{2}$	$p_0(t) = f\left(\frac{a+b}{2}\right) \in \mathbb{R}_0[t]$	$(b-a)f\left(\frac{a+b}{2}\right)$
Trapèze	$a, b$	$p_1(t) = \frac{f(b)-f(a)}{b-a}(t-a) + f(a) \in \mathbb{R}_1[t]$	$\frac{(b-a)}{2}(f(a) + f(b))$

## 4.2 EDO : calcul analytique vs approximation numérique

Les équations différentielles décrivent l'évolution de nombreux phénomènes dans des domaines variés. Une équation différentielle est une équation impliquant une ou plusieurs dérivées d'une fonction inconnue. Si toutes les dérivées sont prises par rapport à une seule variable, on parle d'équation différentielle ordinaire (EDO). Une équation mettant en jeu des dérivées partielles est appelée équation aux dérivées partielles (EDP).

Une EDO est une équation exprimée sous la forme d'une relation

$$F(y(t), y'(t), y''(t), \dots, y^{(p)}(t)) = g(t)$$

- ★ dont les inconnues sont une **fonction**  $y: I \subset \mathbb{R} \rightarrow \mathbb{R}$  et son **intervalle de définition**  $I$
- ★ dans laquelle cohabitent à la fois  $y$  et ses dérivées  $y', y'', \dots, y^{(p)}$  ( $p$  est appelé l'**ordre** de l'équation).

Si la fonction  $g$ , appelée «second membre» de l'équation, est nulle, on dit que l'équation en question est **homogène**. **Résoudre une équation différentielle**, c'est chercher toutes les fonctions, définies sur un intervalle  $I \subset \mathbb{R}$ , qui satisfont l'équation (on dit aussi intégrer l'équation différentielle).

### 💡 EXEMPLE

Résoudre l'équation différentielle  $y'(t) = -y(t)$  signifie chercher toutes les fonctions

$$\begin{aligned} y: I &\subset \mathbb{R} \rightarrow \mathbb{R} \\ t &\mapsto y = f(t) \end{aligned}$$

telles que  $f'(t) = -f(t)$  pour tout  $t \in I$ . On peut vérifier que  $y(t) = ce^{-t}$  pour tout  $t \in \mathbb{R}$  (où  $c$  est constante réelle quelconque) est une solution de l'EDO (en particulier, pour  $c = 0$  on trouve la solution nulle).

### 💡 Définition 4.1 (Solution générale, solution particulière)

Par le terme *solution générale* d'une EDO on désigne un représentant de l'ensemble des solutions. L'une des solutions de l'EDO sera appelée *solution particulière*. On appelle *courbes intégrales* d'une EDO les courbes représentatives des solutions de l'équation.

Une EDO admet généralement une infinité de solutions. Pour choisir, entre les différentes solutions, celle qui décrit le problème physique, il faut considérer d'autres données qui dépendent de la nature du problème, par exemple la valeur prise par la solution et/ou éventuellement ses dérivées en un ou plusieurs points de l'intervalle d'intégration.

### 💡 EXEMPLE (DYNAMIQUE DES POPULATIONS)

Considérons une population de bactéries dans un environnement confiné dans lequel pas plus de  $B > 0$  individus ne peuvent coexister. On suppose qu'au temps initial le nombre d'individus est égal à  $y_0 \ll B$  et que le taux de croissance des bactéries est une constante positive  $C$ . Alors, la vitesse de croissance de la population est proportionnelle au nombre de bactéries, sous la contrainte que ce nombre ne peut dépasser  $B$ . Ceci se traduit par l'équation différentielle suivante

$$y'(t) = Cy(t) \left(1 - \frac{y(t)}{B}\right) \quad (4.1)$$

dont la solution  $y = y(t)$  représente le nombre de bactéries au temps  $t$ .

L'équation (4.1) admet la famille de solutions

$$y(t) = B \frac{e^{Ct+K}}{1 + e^{Ct+K}}$$

$K$  étant une constante arbitraire. Si on impose la condition  $y(0) = 1$ , on sélectionne l'unique solution correspondant à la valeur  $K = -\ln(B-1)$ .

### 💡 Définition 4.2 (Condition initiale)

Soit une EDO d'ordre  $p$ . Une condition initiale (CI) est un ensemble de relations du type  $y(t_0) = y_0, y'(t_0) = y'_0, \dots, y^{(p-1)}(t_0) = y_0^{(p-1)}$  qui imposent en  $t_0$  les valeurs  $y_0, y'_0, \dots, y_0^{(p-1)}$  respectivement de la fonction inconnue et de ses dérivées jusqu'à l'ordre  $p-1$ .

En pratique, se donner une CI revient à se donner le point  $(t_0, y_0)$  par lequel doit passer le graphe de la fonction solution et la valeur de ses dérivées en ce même point.

### 💡 Définition 4.3 (Problème de CAUCHY)

Soit  $I \subset \mathbb{R}$  un intervalle,  $t_0$  un point de  $I$ ,  $\varphi: I \times \mathbb{R} \rightarrow \mathbb{R}$  une fonction donnée continue par rapport aux deux variables et  $y'$  la dérivée de  $y$  par rapport à  $t$ . On appelle *problème de CAUCHY* le problème

trouver une fonction réelle  $y \in \mathcal{C}^1(I)$  telle que

$$\begin{cases} y'(t) = \varphi(t, y(t)), & \forall t \in I, \\ y(t_0) = y_0, \end{cases} \quad (4.2)$$

avec  $y_0$  une valeur donnée appelée *donnée initiale*.

Si  $\varphi$  ne dépend pas explicitement de  $t$  (*i.e.* si  $\varphi(t, y(t)) = \varphi(y(t))$ ), l'EDO est dite *autonome*.

Résoudre un problème de CAUCHY, c'est chercher toutes les fonctions, définies sur un intervalle  $I \subset \mathbb{R}$ , qui satisfont l'équation et qui vérifient la condition initiale. On aura donc des questions naturelles telles

- ★ trouver toutes les fonctions solutions de l'EDO,
- ★ parmi toutes ces fonctions, choisir celles qui vérifient la CI (existence ? unicité ?),
- ★ parmi toutes ces fonctions, étudier le domaine de définition (pour chaque fonction trouvée, quel est le plus grande domaine de définition qui contient le point  $t_0$ ?)

#### Proposition 4.4

Le problème de Cauchy (4.2) est équivalent à l'équation intégrale

$$y(t) = y_0 + \int_{t_0}^t \varphi(s, y(s)) \, ds. \quad (4.3)$$

#### PREUVE

En intégrant l'EDO entre  $t_0$  et  $t$  et en considérant la donnée initiale (4.2) on obtient

$$y(t) = y_0 + \int_{t_0}^t \varphi(s, y(s)) \, ds.$$

La solution du problème de Cauchy est donc de classe  $\mathcal{C}^1(I)$  sur  $I$  et satisfait l'équation intégrale (4.3).

Inversement, si  $y$  est définie par (4.3), alors elle est continue sur  $I$  et  $y(t_0) = y_0$ . De plus, en tant que primitive de la fonction continue  $\varphi(\cdot, y(\cdot))$ , la fonction  $y$  est de classe  $\mathcal{C}^1(I)$  et satisfait l'équation différentielle  $y'(t) = \varphi(t, y(t))$ .

Ainsi, si  $\varphi$  est continue, le problème de Cauchy (4.2) est équivalent à l'équation intégrale (4.3).

Nous verrons plus loin comment tirer parti de cette équivalence pour les méthodes numériques.

#### EXEMPLE (EXISTENCE ET UNICITÉ SUR $\mathbb{R}$ DE LA SOLUTION D'UN PROBLÈME DE CAUCHY)

On se donne  $\varphi(t, y(t)) = 3t - 3y(t)$  et  $y_0 = \alpha$  (un nombre quelconque). On cherche une fonction  $y: t \in \mathbb{R} \mapsto y(t) \in \mathbb{R}$  qui satisfait

$$\begin{cases} y'(t) = 3t - 3y(t), & \forall t \in \mathbb{R}, \\ y(0) = \alpha. \end{cases}$$

Sa solution, définie sur  $\mathbb{R}$ , est donnée par  $y(t) = (\alpha + 1/3)e^{-3t} + t - 1/3$ . En effet on a bien

$$y(0) = (\alpha + 1/3)e^0 + 0 - 1/3 = \alpha, \quad y'(t) = -3(\alpha + 1/3)e^{-3t} + 1 = -3(\alpha + 1/3)e^{-3t} + 1 - 3t + 3t = -3y(t) + 3t.$$

Cet exemple montre le cas où il existe une et une seule solution du problème de CAUCHY définie sur  $\mathbb{R}$ . Les choses ne se passent pas toujours si bien. Les exemples ci-dessous montrent que l'étude mathématique de l'existence et de l'unicité des solutions d'un problème de CAUCHY peut être une affaire délicate.

#### EXEMPLE (NON UNICITÉ DE LA SOLUTION D'UN PROBLÈME DE CAUCHY)

On se donne  $\varphi(t, y(t)) = \sqrt[3]{y(t)}$  et  $y_0 = 0$ . On cherche une fonction  $y: t \in \mathbb{R}^+ \mapsto y(t) \in \mathbb{R}$  qui satisfait

$$\begin{cases} y'(t) = \sqrt[3]{y(t)}, & \forall t > 0, \\ y(0) = 0. \end{cases}$$

On vérifie que les fonctions  $y_1(t) = 0$  et  $y_{2,3}(t) = \pm\sqrt[3]{8t^3/27}$ , pour tout  $t \geq 0$ , sont toutes les trois solution du problème de CAUCHY donné. Cet exemple montre qu'*un problème de CAUCHY n'a pas nécessairement de solution unique*.

EXEMPLE (NON UNICITÉ DE LA SOLUTION D'UN PROBLÈME DE CAUCHY)

On se donne  $\varphi(t, y(t)) = |y(t)|^\alpha$  avec  $\alpha \in ]0; 1[$  et  $y_0 = 0$ . On cherche une fonction  $y: t \in \mathbb{R}^+ \mapsto y(t) \in \mathbb{R}$  qui satisfait

$$\begin{cases} y'(t) = |\varphi(t, y(t))|^\alpha, & \forall t > 0, \\ y(0) = 0. \end{cases}$$

On vérifie que, pour tout  $c \in \mathbb{R}^+$ , les fonctions

$$y_c(t) = \begin{cases} (1-\alpha)^{1/(1-\alpha)}(x-c)^{1/(1-\alpha)} & \text{si } x \geq c, \\ 0 & \text{si } 0 \leq x \leq c \end{cases}$$

sont solution du problème de CAUCHY donné.

Cet exemple montre qu'*un problème de CAUCHY peut admettre une infinité de solutions*.

Notons que pour  $\alpha \geq 1$  le problème de CAUCHY donné admet une et une seule solution, la fonction  $y(t) = 0$  pour tout  $t \in \mathbb{R}^+$ .

EXEMPLE (EXISTENCE ET UNICITÉ SUR  $I \subset \mathbb{R}$  (MAIS NON EXISTENCE SUR  $\mathbb{R}$ ) DE LA SOLUTION D'UN PROBLÈME DE CAUCHY)

On se donne  $\varphi(t, y(t)) = (y(t))^3$  et  $y_0 = 1$ . On cherche une fonction  $y: t \in \mathbb{R}^+ \mapsto y(t) \in \mathbb{R}$  qui satisfait

$$\begin{cases} y'(t) = (y(t))^3, & \forall t > 0, \\ y(0) = 1. \end{cases}$$

On vérifie que la solution  $y$  est donnée par  $y(t) = \frac{1}{\sqrt[3]{1-2t}}$  qui n'est définie que pour  $t \in [0; 1/2[$ . Cet exemple montre qu'*un problème de CAUCHY n'a pas toujours une solution pour tout  $t \in [0; +\infty[$*  puisqu'ici la solution explose lorsque  $t$  tend vers la valeur  $1/2$  (en effet, nous avons  $\lim_{t \rightarrow (1/2)^-} y(t) = +\infty$ ) : le graphe de la solution a une asymptote verticale en  $t = 1/2$ . On parle d'*explosion de la solution en temps fini* ou encore de *barrière*.

Ceci est un phénomène général : pour une solution d'une EDO, la seule façon de ne pas être définie sur  $\mathbb{R}$  est d'avoir un asymptote verticale.

De façon générale, lorsqu'on se donne une équation différentielle et une condition initiale  $y(t_0) = y_0$ , on cherche un intervalle  $I$ , contenant  $t_0$ , sur lequel une solution existe, et qui soit «le plus grand possible» : il n'existe pas d'intervalle plus grand sur lequel l'équation différentielle ait une solution.

Dans ce cours, nous nous contentons de rappeler un résultat d'existence et d'unicité local, au sens où on peut intégrer le problème de CAUCHY jusqu'à  $t < \infty$ .

**Théorème 4.5 (de CAUCHY-LIPSCHITZ, Existence locale et unicité des solutions)**

Considérons une fonction

$$\begin{aligned} \varphi: I \times J &\rightarrow \mathbb{R} \\ (t, y) &\mapsto \varphi(t, y) \end{aligned}$$

définie pour tout  $t$  dans un intervalle  $I \subset \mathbb{R}$  et pour tout  $y$  dans un intervalle  $J \subset \mathbb{R}$ . On suppose que  $\varphi$  est

- \* continue par rapport à ses deux variables,
- \* uniformément lipschitzienne par rapport à sa deuxième variable, ce qui signifie qu'il existe une constante  $L > 0$  (appelée constante de LIPSCHITZ) telle que

$$|\varphi(t, y_1) - \varphi(t, y_2)| \leq L |y_1 - y_2| \quad \forall t \in I, \forall y_1, y_2 \in J.$$

Alors pour toute CI  $y(t_0) = y_0$  avec  $t_0 \in I$  et  $y_0 \in J$  il existe un intervalle  $U \subset I$  qui contient  $t_0$  et une unique fonction  $y = y(t)$  définie sur  $U$  solution maximale de l'EDO  $y'(t) = \varphi(t, y(t))$  et cette solution est de classe  $\mathcal{C}^1(U)$ .

**Remarque**

Si  $\partial_y \varphi$  est continue alors  $\varphi$  est uniformément lipschitzienne par rapport à  $y$ . En effet, il suffit de prendre pour  $L$  le maximum de  $|\partial_y \varphi|$  sur  $\overline{I \times J}$ .

 **Corollaire 4.6 (de CAUCHY-LIPSCHITZ — cas linéaire)**

Soient  $b$  et  $g$  deux fonctions continues d'un intervalle  $I$  dans  $\mathbb{R}$  et considérons l'équation différentielle

$$y'(x) + b(x)y(x) = g(x).$$

Si  $(x_0, y_0) \in I \times \mathbb{R}$ , il existe une unique solution  $y$  de l'équation différentielle telle que  $y(x_0) = y_0$ .

 **ATTENTION**

La fonction  $\varphi$  (ainsi que sa dérivée partielle  $\partial_y \varphi$ ) est une fonction de deux variables donc vérifier que  $\varphi$  est continue signifie utiliser la notion de limite en deux variables. Les limites unilatérales (*i.e.* de la gauche et de la droite) perdent leur sens et sont remplacées par les nombreuses limites directionnelles possibles. En effet, dès que le domaine se situe dans un espace à deux dimensions au moins, les chemins qui mènent à un point donné peuvent suivre divers trajectoires. Ainsi, l'ensemble des points en lesquels une limite peut être considérée, doit être défini en tenant compte de toutes les possibilités d'accès. Cependant, les fonctions élémentaires telles que les polynômes, les fonctions exponentielles, logarithmiques et trigonométriques sont continues dans leurs domaines de définition respectifs. La continuité des autres fonctions s'établit, le cas échéant, en tant que somme, produit, composée, le quotient (lorsque le dénominateur ne s'annule pas) etc., de fonctions continues. Voici quelques exemples :

- 1.  $f(x, y) = x^2 + y^2 - xy + y$  est continue dans  $\mathbb{R}^2$  (polynôme du second degré à deux variables).
- 2.  $f(x, y, z) = e^y + xy^2$  est continue dans  $\mathbb{R}^3$  (somme d'une exponentielle et d'un polynôme).
- 3.  $f(x, y) = \ln(x + y^2) - 3$  est continue dans  $\{(x, y) \in \mathbb{R}^2 \mid x + y^2 > 0\}$  comme somme du logarithme d'un polynôme (fonction composée) et d'une constante.

 **Théorème 4.7**

Soit  $y = f(t)$  une solution maximale définie sur un intervalle de vie  $U = ]a; b[$ . Si  $b \neq +\infty$  alors  $\lim_{t \rightarrow b^-} y(t) = \infty$ , *i.e.* le graphe de la solution a une asymptote verticale en  $t = b$ . Même chose si  $a \neq -\infty$ .

On utilise souvent le théorème sous forme contraposée : si les solutions ne peuvent pas «exploser», alors elles sont définies sur  $\mathbb{R}$ .

 **Remarque**

D'un point de vue pratique, cet énoncé nous aidera à faire des dessins, en garantissant que les graphes des solutions ne se rencontrent jamais. On peut en déduire quelques remarques plus subtiles :

- \* si l'EDO admet comme solution la solution nulle mais  $y_0 \neq 0$ , alors la solution du problème de CAUCHY est du signe de  $y_0$  pour tout  $t \in I$ ;
- \* si l'EDO admet deux solutions constantes  $y(t) = \kappa_1$  et  $y(t) = \kappa_2$  pour tout  $t \in U$  et  $y_0 \in ]\kappa_1; \kappa_2[$ , alors la solution du problème de CAUCHY vérifie  $y(t) \in ]\kappa_1; \kappa_2[$  pour tout  $t \in \mathbb{R}$ .

*L'hypothèse  $\varphi$  lipschitzienne en  $y$  est très contraignante. Par exemple, considérons la fonction  $\varphi(t, y) = y^2$  et  $I = J = \mathbb{R}$ . Elle est lipschitzienne par rapport à  $y$  sur un intervalle  $U \subset I$  borné mais n'est pas lipschitzienne sur  $\mathbb{R}$ . Elle admet pour solutions maximales :*

1. la fonction  $t \mapsto y(t) = 0$  pour tout  $t \in \mathbb{R}$ , qui est globale,
2. les fonctions  $t \mapsto y(t) = 1/(c - t)$  pour tout  $t \in \mathbb{R} \setminus \{c\}$ .

*Parmi les solutions maximales, une seule est donc globale. On vérifie en effet que les solutions non nulles de  $y'(t) = y^2(t)$  tendent vers  $+\infty$  en temps fini.*

En pratique, on ne peut expliciter les solutions analytiques que pour des équations différentielles ordinaires très particulières. Dans certains cas, on ne peut exprimer la solution que sous forme implicite.

 **EXEMPLE**

C'est le cas par exemple de l'EDO  $y'(t) = \frac{y(t)-t}{y(t)+t}$  dont les solutions vérifient la relation implicite

$$\frac{1}{2} \ln(t^2 + y^2(t)) + \arctan\left(\frac{y(t)}{t}\right) = C,$$

où  $C$  est une constante arbitraire.

Dans d'autres cas, on ne parvient même pas à représenter la solution sous forme implicite.

### EXEMPLE

C'est le cas par exemple de l'EDO  $y'(t) = e^{-t^2}$  dont les solutions ne peuvent pas s'écrire comme composition de fonctions élémentaires.

Pour ces raisons, on cherche des méthodes numériques capables d'approcher la solution de toutes les équations différentielles qui admettent une solution.

## 4.2.1 EDO d'ordre 1 à variables séparables

Lorsque l'équation est de la forme

$$f(y(x))y'(x) = g(x)$$

où  $f$  et  $g$  sont des fonctions données dont on connaît des primitives  $F$  et  $G$ , on a

$$F(y(x)) = G(x) + C \quad \text{où } C \in \mathbb{R},$$

et si  $F$  possède une fonction réciproque  $F^{-1}$ , on en déduit

$$y(x) = F^{-1}(G(x) + C),$$

relation qui donne toutes les solutions de l'équation. Cette solution générale dépend de la constante d'intégration  $C$ .

### Astuce (Astuce mnémotechnique)

En pratique, étant donné que  $y'(x) = dy/dx$ , on peut écrire l'équation  $f(y(x))y'(x) = g(x)$  sous la forme

$$f(y) dy = g(x) dx,$$

puis intégrer formellement les deux membres

$$\int f(y) dy = \int g(x) dx,$$

pour obtenir  $F(y) = G(x) + C$  et exprimer  $y$  en fonction de  $x$ .

### EXEMPLE

On veut résoudre l'équation différentielle  $y'(x) = xy(x)$  sur des intervalles à préciser. Il s'agit d'une EDO du premier ordre à variables séparables :

- \* *Recherche des solutions constantes.* Si  $y(x) = A$  pour tout  $x$  alors  $y'(x) = 0$  pour tout  $x$  et l'EDO devient  $0 = xA$  pour tout  $x$ . Par conséquent  $A = 0$  : la fonction  $y(x) = 0$  pour tout  $x$  est l'unique solution constante de l'EDO.
- \* *Recherche des solutions non constantes.* La fonction  $y(x) = 0$  pour tout  $x$  étant solution, toute autre solution  $x \mapsto y(x)$  sera donc non nulle. On peut alors diviser l'EDO par  $y$  et procéder formellement comme suit :

$$\frac{y'(x)}{y(x)} = x \implies \int \frac{1}{y} dy = \int x dx \implies \ln|y| = \frac{x^2}{2} + C \text{ avec } C \in \mathbb{R}.$$

Ainsi, toute solution non nulle est de la forme

$$y(x) = D e^{x^2/2} \quad \text{avec } D \in \mathbb{R}^*.$$

## 4.2.2 EDO d'ordre 1 linéaire

Elles sont de la forme

$$a(x)y'(x) + b(x)y(x) = g(x)$$

où  $a$ ,  $b$  et  $g$  sont des fonctions données, continues sur un intervalle  $I \subset \mathbb{R}$ . Pour la résolution, on se place sur un intervalle  $J \subset I$  tel que la fonction  $a$  ne s'annule pas sur  $J$ .

Pour  $x \in \mathcal{D}_b \cap \mathcal{D}_g \cap \{x \in \mathcal{D}_a \mid a(x) \neq 0\}$ , toute solution  $y(x)$  de cette EDO peut être écrite soit comme somme de deux fonctions ( $y_H$  et  $y_P$ ) soit comme produit de deux fonctions ( $u$  et  $v$ ) :

$$y(x) = \underbrace{C e^{-A(x)}}_{y_H(x)} + \underbrace{B(x) e^{-A(x)}}_{y_P(x)}$$

avec

- \*  $A(x)$  une primitive de  $\frac{b(x)}{a(x)}$ ,
- \*  $B(x)$  une primitive de  $\frac{g(x)}{a(x)} e^{A(x)}$ .

On peut montrer que

- \*  $y_H$  est la solution générale de l'EDO homogène associée, c'est-à-dire de l'EDO  $a(x)y'(x) + b(x)y(x) = 0$  (qui est à variables séparables);

En effet, la fonction  $y(x) = 0$  pour tout  $x$  étant solution, toute autre solution  $x \mapsto y(x)$  sera donc non nulle. On peut alors diviser l'EDO homogène associée par  $y$  et procéder formellement comme suit :

$$\frac{y'(x)}{y(x)} = -\frac{b(x)}{a(x)} \implies \int \frac{1}{y} dy = -\int \frac{b(x)}{a(x)} dx \implies \ln|y| = -\int \frac{b(x)}{a(x)} dx.$$

Ainsi, toute solution non nulle de l'équation homogène associée est de la forme

$$y_H(x) = Ce^{-A(x)} \quad \text{où } A(x) = \int \frac{b(u)}{a(u)} du$$

avec  $C$  constante arbitraire.

- \*  $y_P$  est une solution particulière.

Cette solution particulière peut être une solution «évidente», par exemple une solution constante. Dans la quête d'une solution évidente (non constante) le principe de superposition peut être utile : soient  $a$  et  $b$  deux réels et  $g_1, g_2, \dots, g_n$   $n$  des applications continues sur un intervalle  $I$  de  $\mathbb{R}$ . Si  $y_k$  est une solution particulière de l'EDO  $ay'(x) + by(x) = g_k(x)$  alors  $\sum_{k=1}^n y_k$  est une solution particulière de l'EDO  $ay'(x) + by(x) = \sum_{k=1}^n g_k(x)$ .

Si on ne trouve pas de solution particulière on peut en chercher une par la méthode de LAGRANGE ou de variation de la constante. Si  $y_1(x)$  est une solution non nulle de l'EDO homogène, on introduit une fonction auxiliaire inconnue  $B(x)$  telle que  $y(x) = B(x)y_1(x)$  soit solution de notre EDO. On calcule alors  $y'(x)$  et on reporte  $y'(x)$  et  $y(x)$  dans notre EDO. On observe que  $K(x)$  disparaît, ce qui fournit une auto-vérification. Il ne reste que  $B'(x)$ , ce qui permet de calculer  $B(x)$  et donc  $y_P(x)$ .

#### EXEMPLE

Considérons l'EDO

$$y'(x) - y(x) = x.$$

On a

$$a(x) = 1, \quad b(x) = -1, \quad g(x) = x.$$

Pour  $x \in \mathbb{R}$  on a

- \*  $A(x) = \int -1 dx = -x,$
- \*  $B(x) = \int xe^{-x} dx = -(1+x)e^{-x},$

donc

$$y(x) = (C - (1+x)e^{-x}) e^x = Ce^x - (1+x).$$

Il faut prendre garde à déterminer le domaine de validité  $\mathcal{D}$  de la solution avant tout calcul. Ce domaine de validité dépend du domaine de continuité des fonctions qui sont coefficients de l'EDO. La fonction solution peut avoir un domaine de définition plus grand, mais elle n'est solution que sur l'intervalle  $\mathcal{D}$ .

#### EXEMPLE (IMPORTANCE DU DOMAINE D'INTÉGRATION)

On se propose de résoudre l'équation différentielle

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

et étudier le domaine de définition de ses solutions.

Il s'agit d'une équation différentielle linéaire du premier ordre. Comme  $a(x) = x$ , on cherche sa solution générale sur  $I_1 = ]-\infty; 0[$  ou sur  $I_2 = ]0, +\infty[$ .

- \* *Résolution de l'équation homogène associée :  $xy'(x) - y(x) = 0$ .*

$y(x) = 0$  pour tout  $x$  étant solution, les autres solutions ne s'annulent ni sur  $I_1$  ni sur  $I_2$ . On peut donc diviser par  $y$  et procéder formellement :  $\frac{y'(x)}{y(x)} = \frac{1}{x}$ , soit  $\int \frac{dy}{y} = \int \frac{dx}{x}$  d'où  $\ln|y(x)| = \ln|x| + D$ , ce qui conduit à la solution générale sur  $I_K$  de l'équation homogène

$$y_H(x) = Cx \quad \text{avec } C \in \mathbb{R}.$$

\* Recherche d'une solution particulière (méthode de LAGRANGE).

Considérons une nouvelle fonction inconnue  $K(x)$  telle que  $y_P(x) = K(x)x$  soit solution de  $xy'(x) - y(x) = \frac{2x+1}{x^2+1}$ . On calcule  $y'_P(x) = K'(x)x + K(x)$  et on le reporte dans l'EDO ; on obtient<sup>1</sup>

$$x[K'(x)x + K(x)] - K(x)x = \frac{2x+1}{x^2+1} \implies K'(x) = \frac{2x+1}{x^2(x^2+1)} = \frac{1}{x^2} + \frac{2}{x} - \frac{2x}{x^2+1} - \frac{1}{x^2+1}.$$

On en déduit

$$K(x) = -\frac{1}{x} + 2\ln|x| - \ln(x^2+1) - \arctan(x)$$

et donc

$$y_P(x) = -1 + 2x\ln|x| - x\ln(x^2+1) - x\arctan(x).$$

La solution générale est donc

$$y(x) = y_H(x) + y_P(x) = -1 + 2x\ln|x| - x\ln(x^2+1) - x\arctan(x) + Cx \quad \text{avec } C \in \mathbb{R}.$$

On a  $\lim_{x \rightarrow 0} y(x) = -1$ . Pour que la fonction définie sur  $\mathbb{R}$  avec un prolongement par continuité en 0 par

$$y(x) = \begin{cases} -1 + 2x\ln|x| - x\ln(x^2+1) - x\arctan(x) + C_1x & \text{si } x < 0, \\ -1 & \text{si } x = 0, \\ -1 + 2x\ln|x| - x\ln(x^2+1) - x\arctan(x) + C_2x & \text{si } x > 0, \end{cases}$$

soit un prolongement de la solution de notre EDO sur  $\mathbb{R}$  il faut qu'elle soit de classe  $C^1(\mathbb{R})$ . Puisque  $\lim_{x \rightarrow 0} \frac{y(x)+1}{x} = -\infty$ , elle n'est pas dérivable en  $x = 0$  et on conclut qu'il n'existe aucune solution de l'EDO définie sur  $\mathbb{R}$ .

EXEMPLE (LOI DE NEWTON

Considérons une tasse de café à la température de  $75^\circ\text{C}$  dans une salle à  $25^\circ\text{C}$ . Après 5 minutes le café est à  $50^\circ\text{C}$ . Si on suppose que la vitesse de refroidissement du café est proportionnelle à la différence des températures (*i.e.* que la température du café suit la loi de Newton), cela signifie qu'il existe une constante  $\gamma < 0$  telle que la température vérifie l'EDO du premier ordre

$$T'(t) = \gamma(T(t) - 25)$$

avec la CI

$$T(5) = 50,$$

ayant convenu qu'une unité de temps correspond à une minute et la température est mesuré en degré Celsius.

1. On commence par calculer toutes les solutions de l'EDO. Étant une équation différentielle du premier ordre, la famille de solutions dépendra d'une constante qu'on fixera en utilisant la CI. Si on réécrit l'EDO sous la forme  $T'(t) - \gamma T(t) = -25\gamma$ , on a une EDO linéaire d'ordre 1 avec  $a(t) = 1$ ,  $b(t) = -\gamma$  et  $g(t) = -25\gamma$ . Donc

$$\begin{aligned} * \ A(t) &= \int -\gamma dt = -\gamma t, \\ * \ B(t) &= \int -25\gamma e^{A(t)} dt = 25 \int e^{-\gamma t} dt = 25e^{-\gamma t}. \end{aligned}$$

Toutes les solutions de l'EDO sont les fonctions  $T(t) = De^{\gamma t} + 25$  pour  $D \in \mathbb{R}$ .

Notons que la seule solution constante est la fonction  $T(t) = 25$  pour tout  $t > 0$ .

2. La valeur numérique de la constante d'intégration  $D$  est obtenue grâce à la CI :

$$75 = T(0) = 25 + De^{\gamma \cdot 0} \implies D = 50 \implies T(t) = 25 + 50e^{\gamma t}.$$

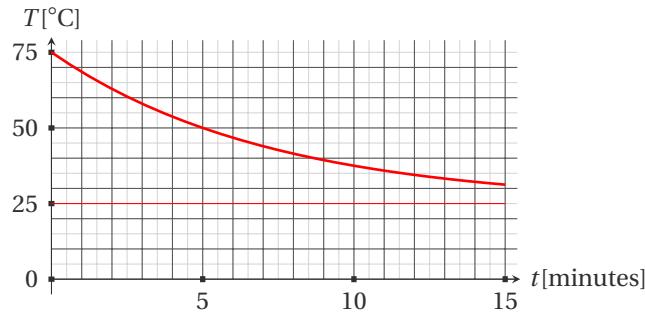
3. Il ne reste qu'à établir la valeur numérique de la constante de refroidissement  $\gamma$  grâce à l'*«indice»* :

$$50 = T(5) = 25 + 50e^{\gamma \cdot 5} \implies \gamma = -\frac{\ln(2)}{5} \implies T(t) = 25 + 50e^{-\frac{\ln(2)}{5}t}$$

On peut donc conclure que la température du café évolue selon la fonction

$$T(t) = 25 + 50e^{-\frac{\ln(2)}{5}t}.$$

1. Intégration de fractions simples :  $\frac{2x+1}{x^2(x^2+1)} = \frac{A}{x} + \frac{B}{x^2} + \frac{Cx+D}{x^2+1} \iff 2x+1 = Ax(x^2+1) + B(x^2+1) + (Cx+D)x^2$ . Si  $x=0$  on obtient que  $B=1$  et donc  $-x+2 = A(x^2+1) + (Cx+D)x$ . Si  $x=0$  on obtient que  $A=2$  et donc  $-2x-1 = Cx+D$  d'où  $D=-1$  et  $C=-2$ .



### 4.2.3 EDO d'ordre 1 de Bernoulli

Elles sont du premier ordre et de la forme

$$u(x)y'(x) + v(x)y(x) = w(x)(y(x))^\alpha, \quad \alpha \in \mathbb{R} \setminus \{0; 1\}$$

où  $u$ ,  $v$  et  $w$  sont des fonctions données, continues sur un intervalle  $I \subset \mathbb{R}$ . Pour la résolution, on se place sur un intervalle  $J \subset I$  tel que la fonction  $u$  ne s'annule pas sur  $J$  et on définit une nouvelle fonction  $x \mapsto z(x) = (y(x))^{1-\alpha}$ . L'EDO initiale est alors équivalente à l'EDO linéaire du premier ordre suivante :<sup>2</sup>

$$\underbrace{u(x)z'(x)}_{a(x)} + \underbrace{(1-\alpha)v(x)z(x)}_{b(x)} = \underbrace{(1-\alpha)w(x)}_{g(x)}.$$

Par conséquent, pour  $x \in \mathcal{D}_v \cap \mathcal{D}_w \cap \{x \in \mathcal{D}_u \mid u(x) \neq 0\}$ , toute solution  $y$  s'écrit comme  $y(x) = [z(x)]^{1/(1-\alpha)}$  avec

- \*  $z(x) = \underbrace{Ce^{-A(x)}}_{y_H(x)} + \underbrace{B(x)e^{-A(x)}}_{y_P(x)}$ ,
- \*  $A(x)$  une primitive de  $(1-\alpha)\frac{v(x)}{u(x)}$ ,
- \*  $B(x)$  une primitive de  $(1-\alpha)\frac{w(x)}{u(x)}e^{A(x)}$ .

#### EXEMPLE

On se propose de résoudre l'équation différentielle

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

Il s'agit d'une équation différentielle de BERNOULLI. Comme  $u(x) = 1$  pour tout  $x \in \mathbb{R}$ , on cherche sa solution générale sur  $\mathbb{R}$ .

- \*  $A(x) = (1-\alpha) \int \frac{v(x)}{u(x)} dx = -2 \int \frac{1/2}{1} dx = -x,$
- \*  $B(x) = (1-\alpha) \int \frac{w(x)}{u(x)} e^{A(x)} dx = -2 \int \frac{(x-1)/2}{1} e^{-x} dx = \int (1-x)e^{-x} dx = -(1-x)e^{-x} - \int e^{-x} dx = xe^{-x},$
- \*  $z(x) = (C + B(x))e^{-A(x)} = (C + xe^{-x})e^x = Ce^x + x,$

et on conclut que la solution générale de l'EDO de BERNOULLI assignée est

$$y(x) = \frac{1}{\sqrt{x + Ce^x}}.$$

Notons que  $y$  n'est définie que si  $x + Ce^x > 0$ .

### 4.2.4 Quelques schémas numériques

Considérons le problème de CAUCHY (4.2) :

trouver une fonction  $y: I \subset \mathbb{R} \rightarrow \mathbb{R}$  définie sur un intervalle  $I$  telle que

$$\begin{cases} y'(t) = \varphi(t, y(t)), & \forall t \in I = ]t_0, T[, \\ y(t_0) = y_0, \end{cases}$$

2. Formellement  $z = y^{1-\alpha}$  implique d'une part  $y = zy^\alpha$  et d'autre part  $z' = (1-\alpha)y^{-\alpha}y'$  et donc  $y' = (1-\alpha)z'y^\alpha$ .

avec  $y_0$  une valeur donnée et supposons que l'on ait montré l'existence et l'unicité d'une solution  $y$  pour  $t \in I$ .

Pour  $h > 0$  soit  $t_n \equiv t_0 + nh$  avec  $n = 0, 1, 2, \dots, N_h$  une suite de nœuds de  $I$  induisant une discrétisation de  $I$  en sous-intervalles  $I_n = [t_n; t_{n+1}]$ . La longueur  $h$  est appelé le *pas de discrétisation*. Le nombre  $N_h$  est le plus grand entier tel que  $t_{N_h} \leq T$ .

Pour chaque nœud  $t_n$ , on cherche la valeur inconnue  $u_n$  qui approche la valeur exacte  $y(t_n)$ . L'ensemble des valeurs  $\{u_0 = y_0, u_1, \dots, u_{N_h}\}$  représente la solution numérique.

Considérons le problème de CAUCHY (4.2) et supposons que l'on ait montré l'existence d'une solution  $y$ . Le principe des méthodes numériques est de subdiviser l'intervalle  $I = [t_0, T]$ , avec  $T < +\infty$ , en  $N_h$  intervalles de longueur  $h = (T - t_0)/N_h = t_{n+1} - t_n$ ;  $h$  est appelé le pas de discrétisation. Si nous intégrons l'EDO  $y'(t) = \varphi(t, y(t))$  entre  $t_n$  et  $t_{n+1}$  nous obtenons

$$y(t_{n+1}) - y(t_n) = \int_{t_n}^{t_{n+1}} \varphi(t, y(t)) dt.$$

Pour chaque nœud  $t_n = t_0 + nh$  ( $1 \leq n \leq N_h$ ) on cherche la valeur inconnue  $u_n$  qui approche  $y(t_n)$ . L'ensemble des valeurs  $\{u_0 = y_0, u_1, \dots, u_{N_h}\}$  représente la solution numérique.

On peut construire différentes schémas selon la formule de quadrature utilisée pour approcher le membre de droite. Les schémas qu'on va construire permettent de calculer (explicitement ou implicitement)  $u_{n+1}$  à partir de  $u_n, u_{n-1}, \dots$  et il est donc possible de calculer successivement  $u_1, u_2, \dots$ , en partant de  $u_0$  par une formule de récurrence de la forme

$$\begin{cases} u_0 = y_0, \\ u_{n+1} = \Phi(u_n, u_{n-1}, \dots, u_{n-k}), \quad \forall n \in \mathbb{N}. \end{cases}$$

#### Définition 4.8 (Méthodes à un pas et méthodes multi-pas)

Une méthode numérique pour l'approximation du problème de CAUCHY (4.2) est dite *à un pas* si pour tout  $n \in \mathbb{N}$ ,  $u_{n+1}$  ne dépend que de  $u_n$ . Autrement, on dit que le schéma est une méthode *multi-pas* (ou à pas multiples).

#### Définition 4.9 (Méthodes explicites et méthodes implicites)

Une méthode est dite *explicite* si la valeur  $u_{n+1}$  peut être calculée directement à l'aide des valeurs précédentes  $u_k$ ,  $k \leq n$  (ou d'une partie d'entre elles). Une méthode est dite *implicite* si  $u_{n+1}$  n'est définie que par une relation implicite faisant intervenir la fonction  $\varphi$ .

### Schémas numériques à un pas

- \* Si on utilise la formule de quadrature du rectangle à gauche, *i.e.*

$$\int_{t_n}^{t_{n+1}} \varphi(t, y(t)) dt \approx h\varphi(t_n, y(t_n))$$

on obtient le *schéma d'EULER progressif*

$$\begin{cases} u_0 = y(t_0) = y_0, \\ u_{n+1} = u_n + h\varphi(t_n, u_n) \quad n = 0, 1, 2, \dots, N_h - 1 \end{cases} \quad (4.4)$$

Il s'agit d'un schéma à 1 pas explicite car il permet d'expliciter  $u_{n+1}$  en fonction de  $u_n$ .

- \* Si on utilise la formule de quadrature du rectangle à droite, *i.e.*

$$\int_{t_n}^{t_{n+1}} \varphi(t, y(t)) dt \approx h\varphi(t_{n+1}, y(t_{n+1}))$$

on obtient le *schéma d'EULER rétrograde*

$$\begin{cases} u_0 = y(t_0) = y_0, \\ u_{n+1} - h\varphi(t_{n+1}, u_{n+1}) = u_n \quad n = 0, 1, 2, \dots, N_h - 1 \end{cases} \quad (4.5)$$

Il s'agit d'un schéma à 1 pas implicite car il ne permet pas d'expliciter directement  $u_{n+1}$  en fonction de  $u_n$  lorsque la fonction  $f$  n'est pas triviale.

- \* Si on utilise la formule de quadrature du point du milieu, *i.e.*

$$\int_{t_n}^{t_{n+1}} \varphi(t, y(t)) dt \approx h\varphi\left(t_n + \frac{h}{2}, y\left(t_n + \frac{h}{2}\right)\right)$$

on obtient un nouveau schéma :

$$\begin{cases} u_0 = y(t_0) = y_0, \\ u_{n+1} = u_n + h\varphi\left(t_n + \frac{h}{2}, u_{n+1/2}\right) \quad n = 0, 1, 2, \dots N_h - 1 \end{cases}$$

où  $u_{n+1/2}$  est une approximation de  $y(t_n + h/2)$ . Nous pouvons utiliser une prédition d'EULER progressive pour approcher le  $u_{n+1/2}$  dans le terme  $\varphi(t_n + h/2, u_{n+1/2})$  par  $\tilde{u}_{n+1/2} = u_n + (h/2)\varphi(t_n, u_n)$ . Nous avons construit ainsi un nouveau schéma appelé **schéma d'EULER modifié** qui s'écrit

$$\begin{cases} u_0 = y(t_0) = y_0, \\ \tilde{u}_{n+1/2} = u_n + (h/2)\varphi(t_n, u_n), \\ u_{n+1} = u_n + h\varphi\left(t_n + \frac{h}{2}, \tilde{u}_{n+1/2}\right) \quad n = 0, 1, 2, \dots N_h - 1 \end{cases} \quad (4.6)$$

Il s'agit d'un schéma à 1 pas explicite car il permet d'expliciter  $u_{n+1}$  en fonction de  $u_n$ .

Si on utilise la formule de quadrature du point milieu sur l'intervalle  $[t_{n-1}; t_{n+1}]$ , i.e.

$$\int_{t_{n-1}}^{t_{n+1}} \varphi(t, y(t)) dt \approx 2h\varphi(t_n, y(t_n))$$

on obtient

$$\begin{cases} u_0 = y(t_0) = y_0, \\ u_1 \approx y(t_1) \\ u_{n+1} = u_{n-1} + 2h\varphi(t_n, u_n) \quad n = 1, 2, \dots N_h - 1 \end{cases}$$

où  $u_1$  est une approximation de  $y(t_1)$ . Nous pouvons utiliser une prédition d'EULER progressive pour approcher  $u_1$ . Nous avons construit ainsi un nouveau schéma appelé **schéma du point milieu** qui s'écrit

$$\begin{cases} u_0 = y(t_0) = y_0, \\ u_1 = u_0 + h\varphi(t_0, u_0), \\ u_{n+1} = u_{n-1} + 2h\varphi(t_n, u_n) \quad n = 1, 2, \dots N_h - 1 \end{cases} \quad (4.7)$$

Il s'agit d'un schéma à 2 pas explicite car il permet d'expliciter  $u_{n+1}$  en fonction de  $u_n$  et de  $u_{n-1}$ .

- \* Si on utilise la formule du trapèze, i.e.

$$\int_{t_n}^{t_{n+1}} \varphi(t, y(t)) dt \approx \frac{h}{2} (\varphi(t_n, y(t_n)) + \varphi(t_{n+1}, y(t_{n+1})))$$

on obtient le **schéma du trapèze ou de CRANK-NICOLSON**

$$\begin{cases} u_0 = y(t_0) = y_0, \\ u_{n+1} - \frac{h}{2}\varphi(t_{n+1}, u_{n+1}) = u_n + \frac{h}{2}\varphi(t_n, u_n) \quad n = 0, 1, 2, \dots N_h - 1 \end{cases} \quad (4.8)$$

Il s'agit à nouveau d'un schéma à 1 pas implicite car il ne permet pas d'expliciter directement  $u_{n+1}$  en fonction de  $u_n$  lorsque la fonction  $f$  n'est pas triviale. En fait, ce schéma fait la moyenne des schémas d'EULER progressif et rétrograde.

- \* Pour éviter le calcul implicite de  $u_{n+1}$  dans le schéma du trapèze, nous pouvons utiliser une prédition d'EULER progressive et remplacer le  $u_{n+1}$  dans le terme  $\varphi(t_{n+1}, u_{n+1})$  par  $\tilde{u}_{n+1} = u_n + h\varphi(t_n, u_n)$ . Nous avons construit ainsi un nouveau schéma appelé **schéma de HEUN**. Plus précisément, la méthode de HEUN s'écrit

$$\begin{cases} u_0 = y(t_0) = y_0, \\ \tilde{u}_{n+1} = u_n + h\varphi(t_n, u_n), \\ u_{n+1} = u_n + \frac{h}{2} (\varphi(t_n, u_n) + \varphi(t_{n+1}, \tilde{u}_{n+1})) \quad n = 0, 1, 2, \dots N_h - 1 \end{cases} \quad (4.9)$$

Il s'agit à nouveau d'un schéma à 1 pas explicite.

### ✿ Remarque

Pour la mise en application d'un schéma il faut aussi prendre en compte l'influence des erreurs d'arrondi. En effet, afin de minimiser l'erreur globale théorique, on pourrait être tenté d'appliquer une méthode avec un pas très petit, par exemple de l'ordre de  $10^{-16}$ , mais ce faisant, outre que le temps de calcul deviendrait irréaliste, très rapidement les erreurs d'arrondi feraient diverger la solution approchée. En pratique il faut prendre  $h$  assez petit pour que la méthode converge assez

rapidement, mais pas trop petit non plus pour que les erreurs d'arrondi ne donnent pas lieu à des résultats incohérent et pour que les calculs puissent être effectués en un temps raisonnable.

### ✿ Remarque (Stabilités des schémas numériques)

De manière générale, un schéma numérique est dit stable s'il permet de contrôler la solution quand on perturbe les données. Il existe de nombreuses notions de stabilité.

Considérons le problème de CAUCHY (4.2) et supposons que l'on ait montré l'existence d'une solution  $y$ . Deux questions naturelles se posent :

- ★ que se passe-t-il lorsqu'on fixe le temps final  $T$  et on fait tendre le pas  $h$  vers 0 ?
- ★ que se passe-t-il lorsqu'on fixe le pas  $h > 0$  mais on fait tendre  $T$  vers l'infini ?

Dans les deux cas le nombre de nœuds tend vers l'infini mais dans le premier cas on s'intéresse à l'erreur en chaque point, dans le deuxième cas il s'agit du comportement asymptotique de la solution et de son approximation.

A première vue, il semble que le schéma d'EULER progressif et le schéma de HEUN soient préférable au schéma d'EULER rétrograde et de CRANK-NICOLSON puisque ces derniers ne sont pas explicites. Cependant, les méthodes d'EULER implicite et de CRANK-NICOLSON sont inconditionnellement A-stables. C'est aussi le cas de nombreuses autres méthodes implicites. Cette propriété rend les méthodes implicites attractives, bien qu'elles soient plus coûteuses que les méthodes explicites.

### ⌚ EXEMPLE (A-STABILITÉ DES MÉTHODES D'EULER EN FONCTION DU PAS)

On considère le problème de CAUCHY

$$\begin{cases} y'(t) = -y(t), \\ y(0) = 1, \end{cases}$$

sur l'intervalle  $[0; 12]$ .

1. Il s'agit d'une EDO à variables séparables. L'unique solution constante de l'EDO est la fonction  $y(t) \equiv 0$ , toutes les autres solutions sont du type  $y(t) = Ce^{-t}$ . Donc l'unique solution du problème de CAUCHY est la fonction  $y(t) = e^{-t}$  définie pour tout  $t \in \mathbb{R}$ .
2. La méthode d'EULER explicite pour cette EDO s'écrit

$$u_{n+1} = (1 - h)u_n.$$

En procédant par récurrence sur  $n$ , on obtient

$$u_{n+1} = (1 - h)^{n+1}.$$

La suite obtenue est une suite géométrique de raison  $q = 1 - h$ . On sait qu'une telle suite

- ★ diverge si  $|q| > 1$  ou  $q = -1$ ,
- ★ est stationnaire si  $q = 1$ ,
- ★ converge vers 0 si  $|q| < 1$ .

De la formule  $u_{n+1} = (1 - h)^{n+1}$  on déduit que

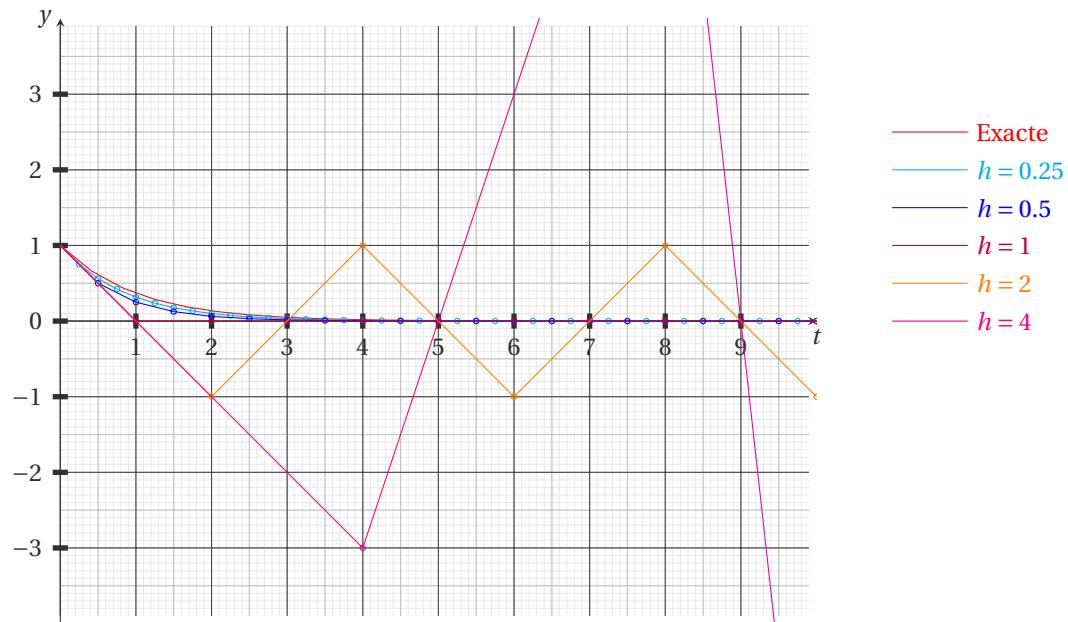
- ★ si  $0 < h < 1$  alors la solution numérique est stable et convergente,
- ★ si  $h = 1$  alors la solution numérique est stationnaire  $u_n = 0$  pour tout  $n \in \mathbb{N}^*$ ,
- ★ si  $1 < h < 2$  alors la solution numérique oscille mais est encore convergente,
- ★ si  $h = 2$  alors la solution numérique oscille, plus précisément on a  $u_{2n} = 1$  et  $u_{2n+1} = -1$  pour tout  $n \in \mathbb{N}^*$ ,
- ★ si  $h > 2$  alors la solution numérique oscille et diverge.

Cela signifie que la méthode est A-stable si et seulement si  $|1 - h| < 1$ .

Voyons ce que cela donne avec différentes valeurs de  $h > 0$  :

- ★ si  $h = 4$  alors  $t_n = 4n$  et  $u_n = (-4)^n$  tandis que  $y(t_n) = e^{-4n}$ ,
- ★ si  $h = 2$  alors  $t_n = 2n$  et  $u_n = (-1)^n$  tandis que  $y(t_n) = e^{-2n}$ ,
- ★ si  $h = 1$  alors  $t_n = n$  et  $u_n = 0$  tandis que  $y(t_n) = e^{-n}$ ,
- ★ si  $h = \frac{1}{2}$  alors  $t_n = n/2$  et  $u_n = \left(\frac{1}{2}\right)^n$  tandis que  $y(t_n) = e^{-n/2}$ ,
- ★ si  $h = \frac{1}{4}$  alors  $t_n = n/4$  et  $u_n = \left(\frac{3}{4}\right)^n$  tandis que  $y(t_n) = e^{-n/4}$ .

Ci-dessous sont tracées sur l'intervalle  $[0; 10]$ , les courbes représentatives de la solution exacte et de la solution calculée par la méthode d'EULER explicite. En faisant varier le pas  $h$  nous pouvons constater que si  $h > 1$  l'erreur commise entre la solution exacte et la solution calculée est amplifiée d'un pas à l'autre.



NB : la première itérée a la même pente quelque soit le pas  $h$  (se rappeler de la construction géométrique de la méthode d'EULER).

3. La méthode d'EULER implicite pour cette EDO s'écrit

$$u_{n+1} = \frac{1}{1+h} u_n.$$

En procédant par récurrence sur  $n$ , on obtient

$$u_{n+1} = \frac{1}{(1+h)^{n+1}}.$$

De la formule  $u_{n+1} = (1+h)^{-(n+1)}$  on déduit que la solution numérique est stable et convergente pour tout  $h$ . En effet, la méthode est inconditionnellement A-stable.

*Remarque :* la suite obtenue est une suite géométrique de raison  $q = 1/(1+h) \in ]0; 1[$ .



## Exercices



### Méthodes des quadratures

#### ★ Exercice 4.1 (Formules de quadrature composites)

On décompose l'intervalle d'intégration  $[a; b]$  en  $m$  sous-intervalles  $[y_j; y_{j+1}]$  tels que  $y_j = a + jH$  où  $H = \frac{b-a}{m}$  pour  $j = 0, 1, \dots, m$ . On utilise alors sur chaque sous-intervalle une formule interpolatoire de nœuds  $\{x_k^{(j)}\}_{k=0}^n$  et de poids  $\{\alpha_k^{(j)}\}_{k=0}^n$  (généralement la même formule sur chaque sous-intervalle). Puisque

$$I_{[a;b]}(f) = \int_a^b f(x) dx = \sum_{j=0}^{m-1} \int_{y_j}^{y_{j+1}} f(x) dx = \sum_{j=0}^{m-1} I_{[y_j;y_{j+1}]}(f),$$

une formule de quadrature interpolatoire composite est obtenue en remplaçant  $I_{[a;b]}(f)$  par

$$\sum_{j=0}^{m-1} \tilde{I}_{[y_j;y_{j+1}]}^{(j)}(f) = \sum_{j=0}^{m-1} \sum_{k=0}^n \alpha_k^{(j)} f(x_k^{(j)})$$

où  $I_{[y_j;y_{j+1}]}(f) \approx \tilde{I}_{[y_j;y_{j+1}]}^{(j)}(f)$ .

Écrire les formules de quadrature composites pour les méthodes des rectangles à gauche, à droite, du point milieu et du trapèze. Pour chaque méthode, écrire une `function` qui prend en entrée `a, b` (avec  $a < b$ ), `f` la fonction à intégrer et `m` ( $\geq 1$ ) le nombre de sous-intervalles et donne en sortie `Int` la valeur de l'intégrale.

#### Correction

**Formule du rectangle à gauche.** La formule du *rectangle à gauche* est obtenue en remplaçant  $f$  par une constante égale à la valeur de  $f$  en la borne gauche de l'intervalle  $[a; b]$  (polynôme qui interpole  $f$  en le point  $(a, f(a))$  et donc de degré 0), ce qui donne

$$\begin{aligned} \tilde{f}(x) &= f(a) \\ I_{[a;b]}(f) &\approx I_{[a;b]}(\tilde{f}) = (b - a)f(a). \end{aligned}$$

On décompose l'intervalle d'intégration  $[a; b]$  en  $m$  sous-intervalles de largeur  $H = \frac{b-a}{m}$  avec  $m \geq 1$ . En introduisant les nœuds de quadrature  $x_k = a + kH$  pour  $k = 0, 1, \dots, m-1$  on obtient la *formule composite du rectangle à gauche*

$$I_{[a;b]}^m(\tilde{f}) = \sum_{k=0}^{m-1} I_{[x_k;x_{k+1}]}(\tilde{f}) = H \sum_{k=0}^{m-1} f(x_k) = H \sum_{k=0}^{m-1} f(a + kH).$$

```
function [Int]=gauche(a,b,f,m)
h=(b-a)/m;
x=[a:h:b];
y=f(x);
Int=h*sum( y(1:end-1) );
end
```

**Formule du rectangle à droite.** La formule du *rectangle à droite* est obtenue en remplaçant  $f$  par une constante égale à la valeur de  $f$  en la borne droite de l'intervalle  $[a; b]$  (polynôme qui interpole  $f$  en le point  $(b, f(b))$  et donc de degré 0), ce qui donne

$$\begin{aligned} \tilde{f}(x) &= f(b) \\ I_{[a;b]}(f) &\approx I_{[a;b]}(\tilde{f}) = (b - a)f(b). \end{aligned}$$

On décompose maintenant l'intervalle d'intégration  $[a; b]$  en  $m$  sous-intervalles de largeur  $H = \frac{b-a}{m}$  avec  $m \geq 1$ . En introduisant les nœuds de quadrature  $x_k = a + (k+1)H$  pour  $k = 0, 1, \dots, m-1$  on obtient la *formule composite du rectangle à droite*

$$I_{[a;b]}^m(\tilde{f}) = \sum_{k=0}^{m-1} I_{[x_k;x_{k+1}]}(\tilde{f}) = H \sum_{k=0}^{m-1} f(x_{k+1}) = H \sum_{k=0}^{m-1} f(a + (k+1)H).$$

```
function [Int]=gauche(a,b,f,m)
h=(b-a)/m;
```

```

x=[a:h:b];
y=f(x);
Int=h*sum( y(2:end) );
end

```

**Formule du rectangle ou du point milieu.** La formule du *rectangle* ou du *point milieu* est obtenue en remplaçant  $f$  par une constante égale à la valeur de  $f$  au milieu de  $[a; b]$  (polynôme qui interpole  $f$  en le point  $\left(\frac{a+b}{2}, f\left(\frac{a+b}{2}\right)\right)$  et donc de degré 0), ce qui donne

$$\tilde{f}(x) = f\left(\frac{a+b}{2}\right)$$

$$I_{[a;b]}(f) \approx I_{[a;b]}(\tilde{f}) = (b-a)f\left(\frac{a+b}{2}\right).$$

On décompose maintenant l'intervalle d'intégration  $[a; b]$  en  $m$  sous-intervalles de largeur  $H = \frac{b-a}{m}$  avec  $m \geq 1$ . En introduisant les noeuds de quadrature  $x_k = a + k \frac{H}{2}$  pour  $k = 0, 1, \dots, 2m$  (*i.e.* chaque sous-intervalle  $[x_{2k}; x_{2k+2}]$  a largeur  $H$  et donc  $x_{2k+1}$  est point milieu), on obtient la *formule composite du point milieu*

$$I_{[a;b]}^m(\tilde{f}) = \sum_{k=0}^{m-1} I_{[x_{2k}; x_{2k+2}]}(\tilde{f}) = \sum_{k=0}^{m-1} (x_{2k+2} - x_{2k}) f(x_{2k+1}) = H \sum_{k=0}^{m-1} f\left(a + (2k+1)\frac{H}{2}\right).$$

```

function [Int]=milieu(a,b,f,m)
h=(b-a)/m;
x=[a+h/2:h:b];
y=f(x);
Int=h*sum( y );
end

```

**Formule du trapèze.** La formule du *trapèze* est obtenue en remplaçant  $f$  par le segment qui relie  $(a, f(a))$  à  $(b, f(b))$  (polynôme qui interpole  $f$  en les points  $(a, f(a))$  et  $(b, f(b))$  et donc de degré 1), ce qui donne

$$\tilde{f}(x) = \frac{f(b)-f(a)}{b-a}(x-a) + f(a)$$

$$I_{[a;b]}(f) \approx I_{[a;b]}(\tilde{f}) = \frac{b-a}{2} (f(a) + f(b)).$$

Pour obtenir la *formule du trapèze composite*, on décompose l'intervalle d'intégration  $[a; b]$  en  $m$  sous-intervalles de largeur  $H = \frac{b-a}{m}$  avec  $m \geq 1$ . En introduisant les noeuds de quadrature  $x_k = a + kH$  pour  $k = 0, 1, \dots, m-1$  on obtient la *formule composite des trapèzes*

$$I_{[a;b]}^m(\tilde{f}) = \sum_{k=0}^{m-1} I_{[x_k; x_{k+1}]}(\tilde{f}) = \sum_{k=0}^{m-1} \frac{x_{k+1} - x_k}{2} (f(x_k) + f(x_{k+1}))$$

$$= \frac{H}{2} \sum_{k=0}^{m-1} (f(x_k) + f(x_{k+1})) = H \left( \frac{1}{2} f(a) + \sum_{k=1}^{m-1} f(a + kH) + \frac{1}{2} f(b) \right).$$

```

function [Int]=trapeze(a,b,f,m)
h=(b-a)/m;
x=[a:h:b];
y=f(x);
Int=0.5*h*( y(1)+2*sum(y(2:end-1))+y(end) );
end

```

## Exercice 4.2

Estimer  $\int_0^{5/2} f(x) dx$  à partir des données

$x$	0	$1/2$	1	$3/2$	2	$5/2$
$f(x)$	$3/2$	2	2	1.6364	1.2500	0.9565

en utilisant

1. la méthode des rectangles à gauche composite,
2. la méthode des rectangles à droite composite,

### 3. la méthode des trapèzes composite.

#### Correction

On a  $a = 0$ ,  $b = \frac{5}{2}$  et  $m = 5$  donc  $h = \frac{b-a}{m} = \frac{1}{2}$ .

Méthode	$\int_a^b f(t) dt \simeq$
Méthode 1	$h \sum_{i=0}^{m-1} f(a + ih) = \frac{1}{2} \left( \frac{3}{2} + 2 + 2 + 1.6364 + 1.2500 \right) = 4.1932$
Méthode 2	$h \sum_{i=0}^{m-1} f(a + (i+1)h) = \frac{1}{2} (2 + 2 + 1.6364 + 1.2500 + 0.9565) = 3.92145$
Méthode 3	$h \left( \frac{1}{2} f(a) + \sum_{i=1}^{m-1} f(a + ih) + \frac{1}{2} f(b) \right) = \frac{1}{2} \left( \frac{3}{4} + 2 + 2 + 1.6364 + 1.2500 + \frac{0.9565}{2} \right) = 4.057325$

### Exercice 4.3

Étant donnée l'égalité

$$\pi = 4 \left( \int_0^{+\infty} e^{-x^2} dx \right)^2 = 4 \left( \int_0^{10} e^{-x^2} dx + \epsilon \right)^2,$$

avec  $0 < \epsilon < 10^{-44}$ , utiliser la méthode des trapèzes composite à 10 intervalles pour estimer la valeur de  $\pi$ .

#### Correction

La méthode des trapèzes composite à  $m$  intervalles pour calculer l'intégrale d'une fonction  $f$  sur l'intervalle  $[a, b]$  s'écrit

$$\int_a^b f(t) dt \simeq h \left( \frac{1}{2} f(a) + \sum_{i=1}^{m-1} f(a + ih) + \frac{1}{2} f(b) \right) \quad \text{avec } h = \frac{b-a}{m}.$$

Ici on a  $f(x) = e^{-x^2}$ ,  $a = 0$ ,  $b = 10$ ,  $m = 10$  d'où  $h = 1$  et on obtient

$$I \simeq \frac{1}{2} + \sum_{i=1}^{10} e^{-i^2} + \frac{1}{2e^{100}} = \frac{1}{2} + \frac{1}{e} + \frac{1}{e^4} + \frac{1}{e^9} + \frac{1}{e^{16}} + \frac{1}{e^{25}} + \frac{1}{e^{36}} + \frac{1}{e^{49}} + \frac{1}{e^{64}} + \frac{1}{e^{81}} + \frac{1}{2e^{100}},$$

ainsi en utilisant la fonction `trapeze(a,b,f,m)` comme suit

```
f=@(x) [exp(-x.^2)];
Int=trapeze(0,10,f,10)
mypi=4*Int.^2
```

on obtient  $\pi \approx 4I^2 = 3.1422$ .

### Exercice 4.4

On considère l'intégrale

$$I = \int_1^2 \frac{1}{x} dx.$$

1. Calculer la valeur exacte de  $I$ .
2. Évaluer numériquement cette intégrale par la méthode des trapèzes avec  $m = 3$  sous-intervalles.
3. Pourquoi la valeur numérique obtenue à la question précédente est-elle supérieure à  $\ln(2)$ ? Est-ce vrai quelque soit  $m$ ? Justifier la réponse. (On pourra s'aider par un dessin.)

#### Correction

1. Une primitive de  $\frac{1}{x}$  est  $F(x) = \ln(x)$ . La valeur exacte est alors  $I = \left[ \ln(x) \right]_{x=1}^{x=2} = \ln(2)$ .
2. La méthode des trapèzes composite à  $m+1$  points pour calculer l'intégrale d'une fonction  $f$  sur l'intervalle  $[a, b]$  s'écrit

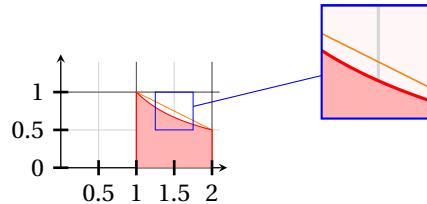
$$\int_a^b f(t) dt \simeq h \left( \frac{1}{2} f(a) + \sum_{i=1}^{m-1} f(a + ih) + \frac{1}{2} f(b) \right) \quad \text{avec } h = \frac{b-a}{m}.$$

Ici on a  $f(x) = \frac{1}{x}$ ,  $a = 1$ ,  $b = 2$ ,  $m = 3$  d'où  $h = \frac{1}{3}$  et on obtient

$$I \simeq \frac{1}{3} \left( \frac{1}{2} f(1) + f(1 + 1/3) + f(1 + 2/3) + \frac{1}{2} f(2) \right) = \frac{1}{3} \left( \frac{1}{2} + \frac{3}{4} + \frac{3}{5} + \frac{1}{4} \right) = \frac{21}{30} = 0,7.$$

```
f=@(x) [1./x];
Int=trapeze(1,2,f,3)
```

3. La valeur numérique obtenue à la question précédente est supérieure à  $\ln(2)$  car la fonction  $f(x) = \frac{1}{x}$  est convexe. On peut se convaincre à l'aide d'un dessin que les trapèzes sont au-dessus de la courbe  $y = 1/x$ , l'aire sous les trapèzes sera donc supérieure à l'aire sous la courbe. Pour bien visualiser la construction considérons  $m = 1$  :



Cela reste vrai quelque soit le pas  $h$  choisi car la fonction est convexe ce qui signifie qu'une corde définie par deux points de la courbe  $y = 1/x$  sera toujours au-dessus de la courbe et par le raisonnement précédent l'aire sous les trapèzes sera supérieure à l'aire exacte.

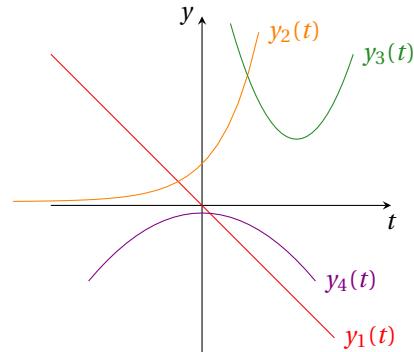
## Étude qualitative d'un problème de Cauchy

### Exercice 4.5

On considère l'équation différentielle

$$y'(t) = \frac{e^t}{t^2 + 1} y(t)$$

Sans résoudre l'équation différentielle, déterminer, parmi les courbes suivantes, celles qui ne représentent sûrement pas une fonction solution de cette EDO et celles qui sont susceptibles d'en représenter une.



### Correction

On remarque que  $\frac{e^t}{t^2+1} > 0$  pour tout  $t \in \mathbb{R}$ ,

Une solution  $y$  de l'EDO doit vérifier  $y'(t) = 0$  si et seulement si  $y(t) = 0$ ; la courbe  $y_1$  (rouge), n'ayant pas de tangente horizontale au point où elle coupe l'axe des abscisses, ne convient donc pas.

L'équation impose aussi que  $y(t)$  et  $y'(t)$  sont de même signe, une condition que ne satisfont pas les courbes  $y_1$  (rouge),  $y_3$  (verte),  $y_4$  (violette).

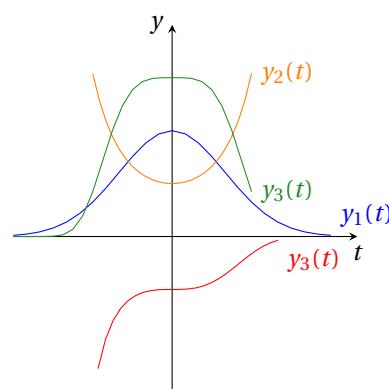
La courbe  $y_2$  (orange) est la seule susceptible de représenter une solution à l'EDO.

### Exercice 4.6

Pour  $t \in \mathbb{R}$ , on considère les quatre équations différentielles

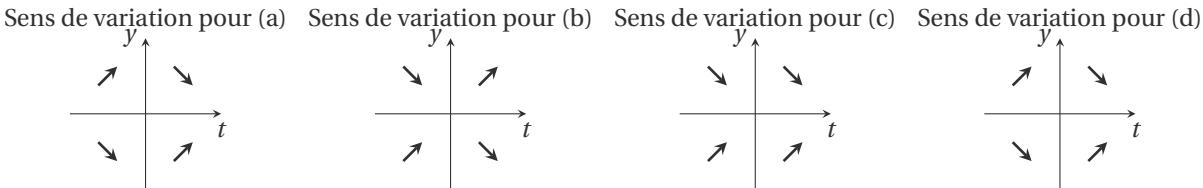
- (a)  $y'(t) = -ty(t)$
- (b)  $y'(t) = ty(t)$
- (c)  $y'(t) = -t^2y(t)$
- (d)  $y'(t) = -t^3y(t)$

Les graphes de ces fonctions sont tracés sur le graphique à coté. Sans résoudre d'équations différentielles, déterminer pour chaque fonction laquelle des courbes suivantes la représente.



**Correction**

Pour chaque EDO on décompose le plan cartésien en quatre partie et on trace le sens de variation de sa solution :



La courbe  $y_3$  (rouge) est la seule où la fonction et sa dérivée sont de signes contraires ; elle ne peut correspondre qu'à la fonction (c). La courbe  $y_2$  (orange) correspond à une fonction ayant même signe que sa dérivée pour  $t > 0$  ; il s'agit donc du graphe de (b). Pour  $t > 1$ , on a  $-t^3 < -t$ , donc le graphe de l'équation (d) est en dessous du graphe de l'équation (a) pour tout  $t > 1$  ; on en déduit que la courbe  $y_1$  (bleue) représente la fonction (a) et que la courbe  $y_4$  (verte) représente la fonction (d).

**Exercice 4.7 (Étude qualitative d'une EDO)**

On considère le problème de Cauchy

$$\begin{cases} y'(t) = 1 - y(t), & t \in \mathbb{R} \\ y(0) = 2. \end{cases}$$

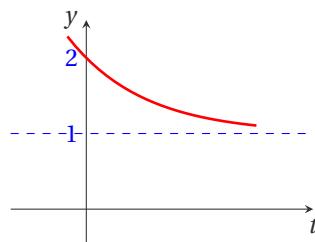
1. Vérifier que le problème satisfait les hypothèses du théorème de CAUCHY-LIPSCHITZ, ce qui permet d'affirmer qu'il admet une et une seule solution  $t \mapsto y(t)$  continue et définie sur  $\mathbb{R}$ ;
2. montrer que la solution est minorée;
3. étudier la monotonie de la solution;
4. calculer la limite pour  $t \rightarrow +\infty$  de la solution;
5. calculer  $y''$  en fonction de  $y$ ;
6. calculer les changements de concavité de la solution;
7. tracer le graphe de la solution.

**Correction**

1.  $\varphi(t, y) = 1 - y$  est continue par rapport à ses deux variables et est uniformément lipschitzienne par rapport à  $y$  car il suffit de prendre  $L = |\partial_y \varphi(t, y)| = 1$  pour tout  $(t, y) \in \mathbb{R}^2$ .
2.  $y(t) = 1$  pour tout  $t \in \mathbb{R}$  est la seule solution constante de l'EDO mais n'est pas solution du problème de Cauchy car  $y(0) \neq 1$ . On sait que la solution du problème de Cauchy est unique, continue, définie sur  $\mathbb{R}$  et passe par le point  $(0, 2)$ , par conséquent

$$y(t) > 1 \quad \forall t \in \mathbb{R}.$$

3. Comme  $y(t) > 1$  pour tout  $t \in \mathbb{R}$ , alors  $y'(t) = 1 - y(t) < 0$  pour tout  $t \in \mathbb{R}$ , ainsi  $y$  est monotone strictement décroissante.
4. La solution est décroissante et minorée donc les limites existent et  $\lim_{t \rightarrow +\infty} y(t) = \ell \geq 1$ . Cela signifie que la droite d'équation  $y = \ell$  est une asymptote horizontale pour le graphe de la solution du problème de Cauchy.
- Si  $\ell > 1$  alors  $\lim_{t \rightarrow +\infty} y'(t) = \lim_{t \rightarrow +\infty} 1 - y(t) = \alpha > 0$ , i.e.  $y$  a une asymptote oblique en  $+\infty$ , ce qui n'est pas possible. Par conséquent  $\lim_{t \rightarrow +\infty} y(t) = 1$ .
5.  $y''(t) = (y'(t))' = (1 - y(t))' = -y'(t) = y(t) - 1$ .
6. Comme  $y(t) > 1$  pour tout  $t \in \mathbb{R}$ , alors  $y''(t) > 0$  pour tout  $t \in \mathbb{R}$  : la solution est convexe.
7. Graphe de la solution :

**Exercice 4.8 (Étude qualitative d'une EDO)**

On considère le problème de CAUCHY

$$\begin{cases} y'(t) = 4 - y^2(t), \\ y(0) = 0. \end{cases}$$

Supposons que le problème admet une et une seule solution  $t \mapsto y(t)$  continue et définie sur  $\mathbb{R}$ .

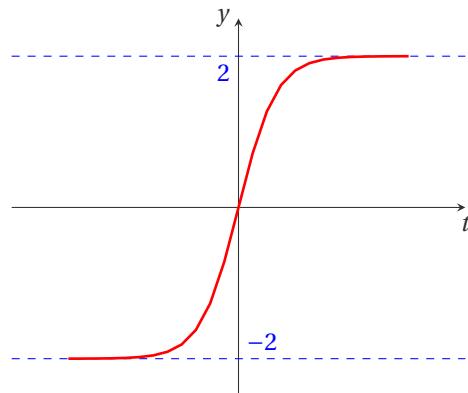
1. Montrer que la solution est bornée et calculer ces bornes ;
2. étudier la monotonie de la solution ;
3. calculer les limites pour  $t \rightarrow \pm\infty$  de la solution ;
4. calculer  $y''$  en fonction de  $y$  ;
5. calculer les changements de concavité de la solution ;
6. tracer le graphe de la solution.

**Correction**

1.  $\varphi(t, y) = 4 - y^2$  est continue par rapport à ses deux variables et est uniformément lipschitzienne par rapport à  $y$  car il suffit de prendre  $L = |\partial_y \varphi(t, y)| = |-2y|$  pour tout  $(t, y) \in \mathbb{R}^2$ .
2. L'EDO se réécrit  $y'(t) = (2 - y(t))(2 + y(t))$ , donc  $y_1(t) = 2$  et  $y_2(t) = -2$  sont deux solutions constantes de l'EDO mais ne sont pas solution du problème de Cauchy car  $y_{1,2}(0) \neq 0$ . On sait que la solution du problème de Cauchy est unique, elle est continue, définie sur  $\mathbb{R}$  et passe par le point  $(0, 0)$ , par conséquent

$$y(t) \in ]-2; 2[ \quad \forall t \in \mathbb{R}.$$

3. Comme  $y(t) \in ]-2; 2[$  pour tout  $t \in \mathbb{R}$ , alors  $y'(t) = 4 - y^2(t) > 0$  pour tout  $t \in \mathbb{R}$ , ainsi  $y$  est monotone strictement croissante.
4. La solution est croissante et bornée donc les limites existent et  $\lim_{t \rightarrow +\infty} y(t) = \ell \leq 2$ . Si  $\ell < 2$  alors  $\lim_{t \rightarrow +\infty} y'(t) = \lim_{t \rightarrow +\infty} 4 - y^2(t) = 4 - \ell^2 > 0$ , i.e.  $y$  a une asymptote oblique en  $+\infty$ , ce qui n'est pas possible car  $y$  est bornée. Par conséquent  $\lim_{t \rightarrow +\infty} y(t) = 2$ .  
Avec le même type de raisonnement on prouve que  $\lim_{t \rightarrow -\infty} y(t) = -2$ .
5.  $y''(t) = (y'(t))' = (4 - y^2(t))' = -2y(t)y'(t) = -2y(t)(4 - y^2(t))$ .
6. Comme  $y(t) \in ]-2; 2[$  pour tout  $t \in \mathbb{R}$ , alors  $y''(t) = 0$  ssi  $y(t) = 0$  et  $y''(t) > 0$  ssi  $y(t) < 0$ ,  $y''(t) < 0$  ssi  $y(t) > 0$ . Comme  $y(t) = 0$  ssi  $t = 0$ , la solution est convexe pour  $t < 0$  et concave pour  $t > 0$ .
7. Graphe de la solution :

**Calcul analytique des solutions d'une EDO d'ordre 1 à variables séparables****Exercice 4.9**

Résoudre le problème de CAUCHY

$$\begin{cases} y'(x) + 2xy^2(x) = 0, \\ y(0) = 2. \end{cases}$$

**Correction**

Il s'agit d'une EDO à variables séparables. La fonction  $y(x) = 0$  pour tout  $x$  est solution de l'EDO mais elle ne vérifie pas la CI. Toute autre solution de l'EDO sera non nulle et se trouve formellement comme suit :

$$y'(x) + 2xy^2(x) = 0 \implies \frac{y'(x)}{y^2(x)} = -2x \implies \int y^{-2} dy = -2 \int x dx \implies y(x) = \frac{1}{x^2 + C}, \quad C \in \mathbb{R}.$$

En imposant la CI on obtient  $2 = 1/C$  d'où l'unique solution du problème de Cauchy :  $y(x) = \frac{2}{x^2 + 1}$ .

**Exercice 4.10**

Résoudre le problème de Cauchy

$$\begin{cases} y'(x) - 4xy^2(x) = 0, \\ y(0) = 2. \end{cases}$$

**Correction**

Il s'agit d'une EDO à variables séparables. La fonction  $y(x) = 0$  pour tout  $x$  est solution de l'EDO mais elle ne vérifie pas la CI. Toute autre solution de l'EDO est non nulle et se trouve formellement comme suit :

$$y'(x) - 4xy^2(x) = 0 \implies \frac{y'(x)}{y^2(x)} = 4x \implies \int y^{-2} dy = 4 \int x dx \implies y(x) = \frac{1}{-2x^2 + c}, c \in \mathbb{R}.$$

En imposant la CI on obtient  $2 = 1/c$  d'où l'unique solution du problème de Cauchy  $y(x) = \frac{2}{1-4x^2}$ .

**Exercice 4.11**

Résoudre le problème de Cauchy

$$\begin{cases} y'(t) = ty^2(t), \\ y(0) = y_0, \end{cases}$$

en fonction de la donnée initiale  $y_0$ .

**Correction**

Il s'agit d'un problème de Cauchy avec une CI  $y(0) = y_0$  et une EDO du premier ordre à variables séparables.

On cherche d'abord les solutions constantes, *i.e.* les fonctions  $y(x) \equiv A \in \mathbb{R}$  qui vérifient l'EDO, c'est-à-dire qui vérifient  $0 = tA^2$  pour tout  $t \in \mathbb{R}$ ; l'unique solution constante est donc la fonction  $y(x) \equiv 0$ .

Comme deux trajectoires ne s'intersectent pas, toutes les autres solution ne s'annulent jamais. Soit donc  $y(x) \neq 0$ ; on peut alors écrire

$$\frac{y'(t)}{y^2(t)} = t \implies \frac{1}{y^2} dy = t dt \implies \int \frac{1}{y^2} dy = \int t dt \implies -\frac{1}{y} = \frac{t^2}{2} + C \implies y(t) = -\frac{1}{\frac{t^2}{2} + C}, \text{ pour } C \in \mathbb{R}.$$

Cette fonction n'est définie que si  $t^2 \neq -2C$ , donc

- ★ si  $C > 0$  alors  $y(t)$  est définie pour tout  $t \in \mathbb{R}$ ,
- ★ si  $C < 0$  alors  $y(t)$  est définie pour tout  $t \in ]-\infty; -\sqrt{-2C}[$  ou  $t \in ]-\sqrt{-2C}; \sqrt{-2C}[$  ou  $t \in ]\sqrt{-2C}; \infty[$ ,
- ★ si  $C = 0$  alors  $y(t)$  est définie pour tout  $t \in ]-\infty; 0[$  ou  $t \in ]0; +\infty[$ .

Comme  $y_0 = y(0) = -\frac{1}{C}$ , la solution du problème de Cauchy est :

- ★ la fonction  $y(t) \equiv 0$  pour tout  $t \in \mathbb{R}$  si  $y_0 = 0$ ;
- ★ la fonction  $y(t) = -\frac{1}{\frac{t^2}{2} - \frac{1}{y_0}}$  pour  $t \in \mathbb{R}$  si  $y_0 < 0$ ;
- ★ la fonction  $y(t) = -\frac{1}{\frac{t^2}{2} - \frac{1}{y_0}}$  pour  $t \in ]-\sqrt{\frac{2}{y_0}}, \sqrt{\frac{2}{y_0}}[$  si  $y_0 > 0$  (c'est-à-dire l'intervalle plus large possible qui contient  $t = 0$ ).

**Exercice 4.12**Soit  $m \in \mathbb{N}^*$ . Montrer que le problème de CAUCHY

$$\begin{cases} y'(t) = y^{2m/(2m+1)}(t), \\ y(0) = 0, \end{cases}$$

admet une infinité de solutions de classe  $\mathcal{C}^1(\mathbb{R})$ . Pourquoi ne peut-on appliquer le théorème de CAUCHY-LIPSCHITZ ?

**Correction**

La solution  $y(t) = 0$  pour tout  $t \in \mathbb{R}$  est une solution du problème donné.

Pour trouver une autre solution commençons par chercher toutes les autres solutions de l'EDO et on imposera ensuite la CI. Il s'agit d'une EDO à variables séparables ainsi, si  $y(t) \neq 0$ , on peut écrire formellement

$$\int y^{-2m/(2m+1)}(t) dy = \int 1 dt$$

d'où la fonction

$$y(t) = \left( \frac{t+c}{2m+1} \right)^{(2m+1)}$$

qui est solution de l'EDO  $y'(t) = y^{2m/(2m+1)}(t)$  pour tout  $c \in \mathbb{R}$ . On vérifie alors aisément que, pour tout  $b \in \mathbb{R}^+$ , les fonctions

$$y_b(t) = \begin{cases} \left( \frac{t+b}{2m+1} \right)^{(2m+1)}, & \text{si } t \leq -b, \\ 0, & \text{si } -b \leq t \leq b, \\ \left( \frac{t-b}{2m+1} \right)^{(2m+1)}, & \text{si } t \geq b, \end{cases}$$

sont de classe  $\mathcal{C}^1(\mathbb{R})$  et sont solution du problème de CAUCHY donné.

En effet, on ne peut pas appliquer le théorème de CAUCHY-LIPSCHITZ car la fonction  $\varphi(t, y) = y^{2m/(2m+1)}$  n'est pas uniformément lipschitzienne par rapport à  $y$  au voisinage de 0 car, pour tout  $y \neq 0$  et pour tout  $L > 0$  on a

$$|\varphi(t, y) - \varphi(t, 0)| = |y^{2m/(2m+1)}| = |y|^{2m/(2m+1)} > L \times |y|.$$

### Exercice 4.13 (Datation au carbone 14)

Le carbone 14 est un isotope présent dans tout organisme vivant. Le nombre d'atomes de carbone 14 est constant tant que l'organisme est en vie. À la mort de l'organisme, le nombre d'atomes décroît avec une vitesse proportionnelle au nombre d'atomes. On note  $n(t) > 0$  le nombre d'atomes au temps  $t$ , exprimé en années, après la mort de l'organisme. Ce mécanisme se traduit par l'équation

$$n'(t) = -kn(t)$$

où  $k$  est une constante positive.

1. Trouver toutes les solutions de l'EDO.
2. Sachant qu'il faut 5700 ans pour que la quantité de carbone 14 diminue de moitié dans un organisme mort, calculer  $k$ .
3. Des ossements anciens récemment exhumés contiennent 9 fois moins de carbone 14 que des ossements similaires d'aujourd'hui. Déterminer l'âge des ossements exhumés.

#### Correction

1. Il s'agit d'une «EDO du premier ordre à variables séparables». Si  $n(t) \equiv c$  est solution alors  $0 = -kc$  d'où  $c = 0$ : l'unique solution constante est la solution  $n(t) = 0$  quelque soit  $t \in \mathbb{R}^+$ .

Si  $n(t) \neq 0$ , on peut écrire

$$\frac{n'(t)}{n(t)} = -k$$

d'où la famille de solutions

$$n(t) = De^{-kt}, \quad D \in \mathbb{R}^+.$$

On conclut que, quelque soit la condition initiale  $n(0) = n_0 \geq 0$ , l'unique solution est  $n(t) = n_0 e^{-kt}$  pour tout  $t \in \mathbb{R}^+$ .

2. Puisque  $n_0/2 = n(5700) = n_0 e^{-5700t}$ , on obtient  $k = \ln 2^{-5700} \approx 1.216 \cdot 10^{-4}$ .

3. Puisque  $n_0/9 = n(\hat{t}) = n_0 e^{-k\hat{t}}$ , on obtient  $\hat{t} = 5700 \frac{\ln 9}{\ln 2} \approx 18000$  ans.

### Exercice 4.14

Deux produits chimiques présents dans une cuve avec une concentration de 1g/l à l'instant  $t = 0$  interagissent et produisent une substance dont la concentration est notée  $y(t)$  à l'instant  $t \geq 0$ . On suppose que  $y(t)$  est régie par l'équation différentielle

$$y'(t) = (1 - y(t))^2 \quad \text{pour tout } t \geq 0.$$

1. Montrer que toute solution de l'EDO est une fonction croissante.
2. Chercher les solutions constantes de l'EDO.
3. Considérons la solution  $y$  telle que  $y(0) = 0$ . Montrer que l'on a  $0 < y(t) < 1$  pour tout  $t > 0$ . (On admettra que les graphes de deux solutions distinctes ne se coupent pas et on pourra s'aider d'un dessin.)
4. Considérons la solution  $y$  telle que  $y(0) = 0$ . Montrer que  $\lim_{t \rightarrow +\infty} y(t) = \ell$  existe. Puis, en admettant que  $\lim_{t \rightarrow +\infty} y'(t) = 0$ , déterminer  $\ell$ .
5. Calculer la solution lorsque  $y(0) = 0$ , lorsque  $y(0) = 1$  et lorsque  $y(0) = 2$ . Dans chacun de ces cas établir l'intervalle maximale d'existence.

**Correction**

1. Pour montrer qu'une fonction est croissante il suffit de montrer que sa dérivée est de signe positif. Si  $y$  est solution de l'EDO on a

$$y'(t) = (1 - y(t))^2 \geq 0 \quad \text{pour tout } t \geq 0$$

car un carré est toujours positif.  $y$  est donc une fonction croissante.

2. On cherche les fonctions constantes solution de l'EDO. Si  $f(t) = c$  est solution de l'EDO alors puisque  $f'(t) = 0$  on obtient

$$0 = (1 - c)^2$$

soit  $c = 1$ . La seule fonction constante solution de l'EDO est la fonction constante égale à 1.

3. Considérons la solution  $y$  telle que  $y(0) = 0$ . Tout d'abord on a montré que la fonction  $y$  était croissante donc  $y(0) \leq y(t)$  pour tout  $t \geq 0$ , par conséquent, puisque  $0 \leq y(0)$ ,  $y(t) \geq 0$  pour tout  $t \geq 0$ . Supposons qu'il existe un  $t_0$  tel que  $y(t_0) \geq 1$ , alors le graphe de  $y$  qui relie continument les points  $(0, y(0))$  et  $(t_0, y(t_0))$  coupe nécessairement le graphe de  $f$ , i.e. la droite d'équation  $y = 1$ . Ceci est impossible, car les graphes de deux solutions distinctes ne se coupent jamais. Il n'existe donc pas de  $t_0$  tel que  $y(t_0) \geq 1$ , c'est-à-dire pour tout  $t \geq 0$ ,  $y(t) < 1$ .

4. Considérons la solution  $y$  telle que  $y(0) = 0$ .

La fonction  $y$  est croissante et majorée par 1, elle admet donc une limite pour  $t \rightarrow +\infty$ . On note  $\lim_{t \rightarrow +\infty} y(t) = \ell$ . On suppose que  $\lim_{t \rightarrow +\infty} y'(t) = \ell$ . En passant à la limite dans l'EDO on obtient :

$$0 = (1 - \ell)^2$$

soit  $\ell = 1$ .

5.  $\star$  Si  $y(0) = 1$  on sait que  $y(t) = 1$  pour tout  $t > 0$ .

$\star$  Si  $y(0) = 0$  on sait que la fonction  $y$  est croissante et  $\lim_{t \rightarrow +\infty} y'(t) = 1$ . En effet, il s'agit d'une EDO à variables séparables et on peut écrire

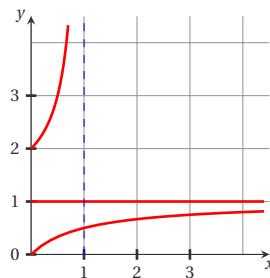
$$\int (1 - y)^{-2} dy = t, \quad \text{i.e.} \quad y(t) = \frac{t + c - 1}{t + c}$$

qui existe sur  $]-\infty; -c[ \cup ] -c; +\infty[$ , d'où, en imposant  $y(0) = 0$ , la solution

$$y(t) = \frac{t}{1+t}, \quad \forall t \geq 0.$$

$\star$  Si  $y(0) = 2$  on sait que la fonction  $y$  est croissante mais elle n'existe que pour  $0 < t < 1$  et on a

$$y(t) = \frac{t-2}{t-1}.$$

**Exercice 4.15 (Logistique)**

Soit  $k$  et  $h$  deux constantes positives. Calculer  $p(t)$  pour  $t > 0$  solution du problème de Cauchy

$$\begin{cases} p'(t) = kp(t) - hp^2(t), \\ p(0) = p_0. \end{cases}$$

Ce modèle, qui décrit l'évolution d'une population de  $p$  individus à l'instant  $t$ , suppose que le taux de croissance du nombre d'individus n'est pas constant mais diminue si la population augmente (les ressources se réduisent).

**Correction**

On doit résoudre l'EDO à variables séparables

$$p'(t) = p(t)(k - hp(t)).$$

On cherche d'abord les solutions constantes, i.e. des fonctions  $p(t) = A$  :

$$0 = A(k - hA) \iff A = 0 \text{ ou } A = \frac{k}{h}.$$

On trouve ainsi deux solutions constantes :

$$p(t) \equiv 0 \quad \text{et} \quad p(t) \equiv \frac{k}{h}.$$

Si on suppose que  $p(t) \neq 0$  et  $p(t) \neq \frac{k}{h}$ , l'EDO se réécrit comme

$$\frac{p'(t)}{p(t)(k - hp(t))} = 1;$$

on doit alors calculer

$$\int \frac{dp}{p(k - hp)} = \int 1 dt$$

i.e.

$$\frac{1}{k} \int \frac{dp}{p} + \int \frac{h}{k - hp} dp = \int 1 dt.$$

On obtient

$$\frac{1}{k} \ln \frac{|p|}{|k - hp|} = (t + C)$$

et en on déduit

$$p(t) = \frac{kDe^{kt}}{1 + hDe^{kt}}.$$

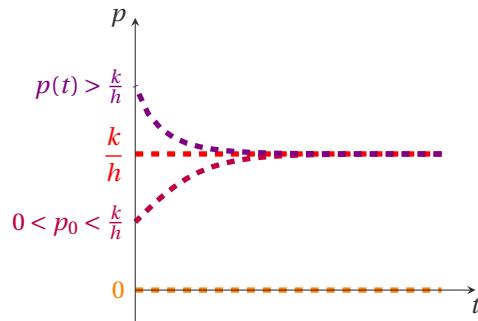
En imposant la condition initiale  $p(0) = p_0$  on trouve la constante d'intégration  $D$  :

$$D = \frac{p_0}{k - hp_0} = \frac{1}{\frac{k}{p_0} - h}.$$

On conclut que toutes les solutions du problème de Cauchy pour  $t \geq 0$  sont

$$p(t) = \begin{cases} 0 & \text{si } p_0 = 0, \\ \frac{k}{h} & \text{si } p_0 = \frac{k}{h}, \\ \frac{1}{\left(\frac{1}{p_0} - \frac{h}{k}\right)e^{-kt} + \frac{h}{k}} & \text{sinon.} \end{cases}$$

Remarquons que  $\lim_{t \rightarrow +\infty} p(t) = \frac{k}{h}$  : une population qui évolue à partir de  $p_0$  individus à l'instant initial selon la loi logistique tend à se stabiliser vers un nombre d'individus d'environ  $k/h$ , ce qui représente la capacité de l'environnement. D'autre part, déjà en analysant l'EDO on aurait pu déduire que les solutions sont des fonctions strictement croissantes si  $p(t) \in ]0, k/h[$ , décroissantes si  $p(t) > k/h$ .

**Exercice 4.16 («Urgence»)**

On étudie la progression d'une maladie contagieuse dans une population donnée. On note  $x(t)$  la proportion des personnes

malades à l'instant  $t$  et  $y(t)$  celle des personnes non atteintes. On a donc  $x(t) + y(t) = 1$  pour tout  $t \geq 0$ . On suppose que la vitesse de propagation de la maladie  $x(t)$  est proportionnelle au produit  $x(t)y(t)$  (ce qui signifie que la maladie se propage par contact). Si on note  $I(t)$  le nombre d'individus infectés à l'instant  $t$  et  $I_T$  le nombre d'individus total, il existe une constante  $k \in \mathbb{R}$  telle que  $I'(t) = kI(t)(I_T - I(t))$ . Si la ville est isolée et compte 5000 individus dont 160 sont malades et 1200 le sont 7 jours après, à partir de quel jour l'infection touchera 80% de la population ? Et 100% ?

### Correction

On a le problème de CAUCHY

$$\begin{cases} I'(t) = kI(t)(5000 - I(t)), & (\text{EDO}) \\ I(0) = 160. & (\text{CI}) \end{cases}$$

Vu la nature de la question on ne s'intéresse qu'aux solutions positives et que pour  $t > 0$ .

1. Tout d'abord on observe qu'il y a deux solutions constantes de l'EDO : la fonction  $I(t) \equiv 0$  et la fonction  $I(t) \equiv 5000$ .
2. Pour chercher toutes les solutions non constantes on remarque qu'il s'agit d'une EDO à variables séparables donc formellement on a

$$\begin{aligned} I'(t) = kI(t)(5000 - I(t)) &\implies \frac{I'(t)}{I(t)(5000 - I(t))} = k &\implies \\ \frac{dI}{I(5000 - I)} = k dt &\implies \int \frac{1}{I(5000 - I)} dI = k \int dt &\implies \\ \int \frac{1}{I} dI - \int \frac{1}{5000 - I} dI = 5000k \int dt &\implies \ln(I) + \ln(5000 - I) = 5000kt + c &\implies \\ \ln \frac{I}{5000 - I} = 5000kt + c &\implies \frac{I}{5000 - I} = D e^{5000kt} &\implies \\ I(t) = \frac{5000 D e^{5000kt}}{1 + D e^{5000kt}} &\implies I(t) = \frac{5000}{D e^{-5000kt} + 1} \end{aligned}$$

3. La valeur numérique de la constante d'intégration  $D$  est obtenue grâce à la CI :

$$160 = I(0) = \frac{5000}{D e^0 + 1} \implies 160 = \frac{5000}{1 + D} \implies D = \frac{4}{121} \implies I(t) = \frac{20000}{4 + 121 e^{-5000kt}}.$$

4. Il ne reste qu'à établir la valeur numérique de la constante  $k$  grâce à l'information sur le nombre d'individus infectés après 7 jours :

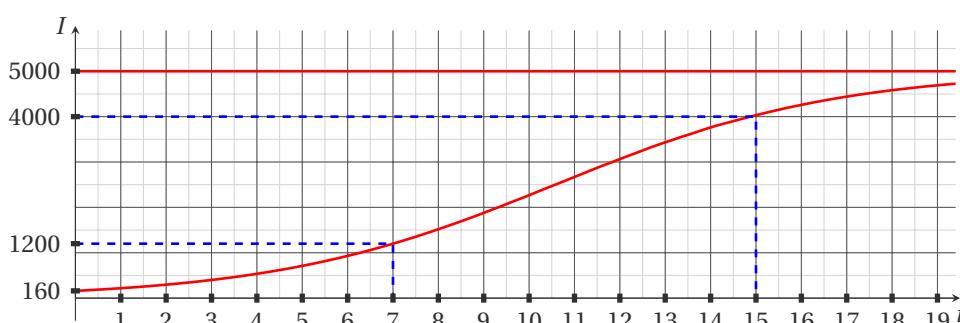
$$1200 = I(7) = \frac{20000}{4 + 121 e^{-35000k}} \implies k = \frac{1}{35000} \ln \frac{363}{38} \implies I(t) = \frac{20000}{4 + 121 e^{-\frac{t}{7} \ln(\frac{363}{38})}}.$$

5. On cherche  $\bar{t}$  tel que  $I(\bar{t}) = 80\%I_T = \frac{80 \times 5000}{100} = 4000$  :

$$4000 = \frac{20000}{4 + 121 e^{-\frac{\bar{t}}{7} \ln(\frac{363}{38})}}$$

d'où  $\bar{t} = \frac{1}{5000} \ln(121) \approx 15$  jours.

6. Avec ce modèle  $\lim_{t \rightarrow +\infty} I(t) = 5000$  mais  $I$  ne peut jamais atteindre exactement 100% de la population en un temps fini (deux solution ne s'intersectent jamais).



### Exercice 4.17

On note  $y(t)$  le nombre de ménages vivant en France équipés d'un ordinateur ( $t$  est exprimé en années et  $y(t)$  en millions de ménages). Le modèle de VARHULST estime que sur la période 1980 – 2020,  $y(t)$  est solution de l'équation différentielle

$$y'(t) = 0,022y(t)(20 - y(t)).$$

1. Calculer toutes les solutions de l'équation différentielle.
2. On pose  $t = 0$  en 1980 et on sait que  $y(0) = 0,01$ . Combien de ménages vivant en France seront équipés d'un ordinateur en 2020 ?

#### Correction

1. On doit résoudre l'EDO à variables séparables

$$y'(t) = 0,022y(t)(20 - y(t)).$$

On cherche d'abord les solutions constantes, *i.e.* des fonctions  $y(t) = A$  pour tout  $t \in \mathbb{R}$  :

$$0 = 0,022A(20 - A) \iff A = 0 \text{ ou } A = 20.$$

On trouve ainsi deux solutions constantes :

$$y(t) \equiv 0 \quad \text{et} \quad y(t) \equiv 20.$$

Si on suppose que  $y(t) \neq 0$  et  $y(t) \neq A$ , l'EDO se réécrit comme

$$\frac{y'(t)}{y(t)(20 - y(t))} = 0,022;$$

on doit alors calculer

$$\int \frac{dy}{y(20 - y)} = \int 0,022 dt,$$

i.e.

$$\frac{1}{20} \left( \int \frac{dy}{y} - \int \frac{1}{y-20} dy \right) = \int 0,022 dt.$$

On obtient

$$\ln \frac{|y|}{|y-20|} = 0,44t + C \quad \text{pour tout } C \in \mathbb{R}$$

et on en déduit

$$y(t) = \frac{20}{1 + 20De^{-0,44t}} \quad \text{pour tout } D \in \mathbb{R}_+.$$

2. Si  $t = 0$  correspond à l'année 1980 et si  $y(0) = 0,01$  alors

$$0,01 = \frac{20}{1 + 20De^{-0,44 \times 0}} \implies D = 1999$$

et la fonction qui estime le nombre de ménages en France équipés d'un ordinateur  $t$  années après 1980 est

$$y(t) = \frac{20}{1 + 1999e^{-0,44t}}.$$

Pour prévoir combien de ménages vivant en France seront équipés d'un ordinateur en 2020 il suffit de calculer  $y(40)$

$$y(40) = \frac{20}{1 + 1999e^{-0,44 \times 40}} \approx 19.99.$$

### Exercice 4.18 (Modèle de GOMPERTZ)

Lorsqu'une nouvelle espèce s'introduit dans un écosystème, elle évolue d'abord lentement; son rythme de croissance s'accélère ensuite à mesure qu'elle s'adapte, puis ralentit quand la population devient trop importante compte tenu des ressources disponibles. Pour ce type d'évolution, on utilise le modèle de GOMPERTZ suivant :

$$y'(t) = -y(t) \ln(y(t)).$$

Calculer toute les solutions de cette équation différentielle pour  $t > 0$  (ne pas oublier les solutions constantes). La population va-t-elle survivre ?

### Correction

- On doit résoudre l'EDO à variables séparables

$$y'(t) = -y(t) \ln(y(t)).$$

On cherche d'abord les solutions constantes, *i.e.* des fonctions  $y(t) = A$  pour tout  $t \in \mathbb{R}$  :

$$0 = A \ln(A) \iff A = 1.$$

On trouve ainsi une solution constante :

$$y(t) \equiv 1.$$

Si on suppose que  $y(t) \neq 1$ , l'EDO se réécrit comme

$$\frac{y'(t)}{y(t) \ln(y(t))} = -1;$$

on doit alors calculer

$$\int \frac{dy}{y \ln(y)} = \int -1 dt.$$

On obtient<sup>3</sup>

$$\ln|\ln(y(t))| = -t + C \quad \text{pour tout } C \in \mathbb{R}$$

et on en déduit

$$y(t) = e^{De^{-t}} \quad \text{pour tout } D \in \mathbb{R}.$$

- Si  $y(0) > 1$  alors  $y'(0) < 0$  (la population décroît) ; si  $0 < y(0) < 1$  alors  $y'(0) > 0$  (la population croît) ; comme  $y(t) = 1$  est solution et comme deux solutions ne peuvent pas se croiser, sans faire de calcul on voit que lorsque  $t$  tend vers l'infini, la population tend vers la valeur d'équilibre  $y(t) = 1$  quelque soit le nombre d'individus à l'instant initial.

## Calcul analytique des solutions d'une EDO d'ordre 1 linéaire

### Exercice 4.19

Résoudre l'équation différentielle

$$(x+1)y'(x) + y(x) = (x+1)\sin(x)$$

sur des intervalles à préciser.

### Correction

L'équation différentielle est linéaire du premier ordre. On la résout sur un intervalle où le coefficient de  $y'(x)$  n'est pas nul, soit sur  $I_1 = ]-\infty; -1[$  ou sur  $I_2 = ]-1; +\infty[$ . Sur chaque intervalle  $I_1$  ou  $I_2$ , l'équation s'écrit

$$[(x+1)y(x)]' = (x+1)\sin(x).$$

En intégrant par parties, on obtient (attention, la constante dépend de l'intervalle)

$$(x+1)y(x) = \int (x+1)\sin(x) dx = -(x+1)\cos(x) + \sin(x) + C.$$

La solution générale sur  $I_1$ , ou sur  $I_2$ , est donc

$$y(x) = -\cos(x) + \frac{\sin(x) + C}{(x+1)} \quad \text{avec } C \in \mathbb{R}.$$

### Exercice 4.20

Résoudre le problème de Cauchy

$$\begin{cases} y'(x) + (3x^2 + 1)y(x) = x^2 e^{-x}, \\ y(0) = 1. \end{cases}$$

<sup>3</sup>  $\int \frac{1}{x \ln(x)} dx = \int \frac{1}{z} dz = \ln|z| + c = \ln|\ln(x)| + C$

**Correction**

On a  $a(x) = 1$ ,  $b(x) = 3x^2 + 1$  et  $g(x) = x^2 e^{-x}$ , donc pour  $x \in \mathbb{R}$  on a

$$\star \quad A(x) = \int \frac{3x^2+1}{1} dx = x^3 + x,$$

$$\star \quad B(x) = \int \frac{x^2 e^{-x}}{1} e^{A(x)} dx = \int x^2 e^{x^3+x} dx = \frac{1}{3} \int 3x^2 e^{x^3+x} dx = \frac{1}{3} \int e^{u(x)} u'(x) dx = \frac{1}{3} e^{u(x)} = \frac{e^{x^3+x}}{3}.$$

Toutes les solutions de l'EDO sont donc les fonctions  $y(x) = \left( C + \frac{e^{x^3+x}}{3} \right) e^{-x^3-x}$  pour  $C \in \mathbb{R}$ .

On cherche parmi ces solutions celle qui vérifie  $y(0) = 1$ ; comme  $y(0) = C + \frac{1}{3}$ , l'unique solution du problème de CAUCHY donné est la fonction  $y(x) = \left( \frac{2}{3} + \frac{e^{x^3+x}}{3} \right) e^{-x^3-x}$ .

**Exercice 4.21**

Résoudre le problème de Cauchy

$$\begin{cases} y'(x) + (3x^2 - 1)y(x) = x^2 e^x, \\ y(0) = -1. \end{cases}$$

**Correction**

On a  $a(x) = 1$ ,  $b(x) = 3x^2 - 1$  et  $g(x) = x^2 e^x$ , donc pour  $x \in \mathbb{R}$  on a

$$\star \quad A(x) = \int \frac{3x^2-1}{1} dx = x^3 - x,$$

$$\star \quad B(x) = \int \frac{x^2 e^x}{1} e^{A(x)} dx = \int x^2 e^{x^3-x} dx = \frac{e^{x^3+x}}{3}.$$

Toutes les solutions de l'EDO sont donc les fonctions  $y(x) = \left( C + \frac{e^{x^3+x}}{3} \right) e^{-x^3+x}$  pour  $C \in \mathbb{R}$ .

On cherche parmi ces solutions celle qui vérifie  $y(0) = -1$ ; comme  $y(0) = C + \frac{1}{3}$ , l'unique solution du problème de CAUCHY donné est la fonction  $y(x) = \left( -\frac{4}{3} + \frac{e^{x^3+x}}{3} \right) e^{-x^3+x}$ .

**Exercice 4.22**

Résoudre le problème de Cauchy

$$\begin{cases} y'(x) + \frac{1}{x-1} y(x) = \frac{(x-2)^2}{x-1}, \\ y(0) = 1. \end{cases}$$

**Correction**

On a  $a(x) = 1$ ,  $b(x) = \frac{1}{x-1}$  et  $g(x) = \frac{(x-2)^2}{x-1}$ .  $b$  est défini pour  $x \neq 1$  et comme on cherche un solution qui passe par le point  $(0, 1)$ , nous allons chercher une solution que pour  $x < 1$ . On a

$$\star \quad A(x) = \int \frac{1}{x-1} dx = \ln(1-x),$$

$$\star \quad B(x) = \int \frac{(x-2)^2}{x-1} e^{A(x)} dx = - \int (x-2)^2 dx = \frac{(x-2)^3}{3}.$$

Toutes les solutions de l'EDO pour  $x < 1$  s'écrivent  $y(x) = \left( C + \frac{(x-2)^3}{3} \right) \frac{1}{x-1}$  pour  $C \in \mathbb{R}$ . On cherche parmi ces solutions celle qui vérifie  $y(0) = 1$ ; comme  $y(0) = -C + \frac{8}{3}$ , l'unique solution du problème de CAUCHY donné est la fonction  $y(x) = \left( \frac{5}{3} + \frac{(x-2)^3}{3} \right) \frac{1}{x-1}$ .

**Exercice 4.23**

Résoudre le problème de Cauchy

$$\begin{cases} y'(x) + (4x^3 + 5)y(x) = x^3 e^{-5x}, \\ y(0) = 1. \end{cases}$$

**Correction**

On a  $a(x) = 1$ ,  $b(x) = 4x^3 + 5$  et  $g(x) = x^3 e^{-5x}$ . On a

$$\star \quad A(x) = \int 4x^3 + 5 dx = x^4 + 5x,$$

$$\star \quad B(x) = \int x^3 e^{-5x} e^{A(x)} dx = - \int x^3 e^{x^4+5x} dx = \frac{e^{x^4}}{4}.$$

Toutes les solutions de l'EDO sont donc les fonctions  $y(x) = \left(C - \frac{e^{-x^4}}{4}\right)e^{-x^4-5x}$  pour  $C \in \mathbb{R}$ .

On cherche parmi ces solutions celle qui vérifie  $y(0) = 1$ ; comme  $y(0) = C + \frac{1}{4}$ , l'unique solution du problème de CAUCHY donné est la fonction  $y(x) = \left(\frac{3}{4} + \frac{e^{-x^4}}{4}\right)e^{-x^4-5x}$ .

### Exercice 4.24

Établir s'il existe des solutions de  $y'(x) = -2y(x) + e^{-2x}$  qui ont dérivée nulle en  $x = 0$ .

#### Correction

On a  $a(x) = 1$ ,  $b(x) = 2$  et  $g(x) = e^{-2x}$ . On a

- \*  $A(x) = \int 2 \, dx = 2x,$
- \*  $B(x) = \int e^{-2x} e^{A(x)} \, dx = \int 1 \, dx = x.$

Toutes les solutions de l'EDO sont donc les fonctions  $y(x) = (C + x)e^{-2x}$  pour  $C \in \mathbb{R}$ .

On cherche si parmi ces solutions il en existe qui vérifient  $y'(0) = 0$ ; comme  $y'(x) = (1 - 2C - 2x)e^{-2x}$  et  $y'(0) = 1 - 2C$ , l'unique solution de l'EDO qui a dérivée nulle en  $x = 0$  est la fonction  $y(x) = (\frac{1}{2} + x)e^{-2x}$ .

### Exercice 4.25

Établir s'il existe des solutions de  $y'(x) = -2xy(x) + x$ .

#### Correction

On a  $a(x) = 1$ ,  $b(x) = 2x$  et  $g(x) = x$ . La solution de cette EDO est du type  $y(x) = y_H(x) + y_P(x)$  où  $y_H(x)$  est la famille de solutions de l'EDO homogène  $y'(x) = -2xy(x)$  et  $y_P(x)$  est une solution particulière de l'EDO complète  $y'(x) = -2xy(x) + x$ . On a  $y_H(x) = Ce^{-A(x)}$  et, par exemple, on cherche  $y_P$  sous la forme  $y_P(x) = K(x)e^{-A(x)}$  avec

- \*  $A(x) = \int \frac{b(x)}{a(x)} \, dx = \int 2x \, dx = x^2,$
- \*  $B(x) = \int \frac{g(x)}{a(x)} e^{A(x)} \, dx = \int xe^{A(x)} \, dx = \int xe^{x^2} \, dx = \frac{1}{2}e^{x^2},$

donc toutes les solutions de l'EDO sont les fonctions  $y(x) = Ce^{-x^2} + \frac{1}{2}$  pour  $C \in \mathbb{R}$ .

Notons qu'il n'est même pas nécessaire de calculer  $B(x)$ ; en effet, il suffit de trouver une solution particulière évidente, par exemple une solution constante. Si  $y(x) = \alpha$  pour tout  $x$  est une solution de l'EDO complète, alors  $0 = -2x\alpha + x$ , i.e.  $\alpha = 1/2$ .

On pose alors  $y_P(x) = 1/2$  et on a  $y(x) = y_H(x) + y_P(x) = Ce^{-x^2} + \frac{1}{2}$ .

Toutes les solutions de l'EDO sont donc les fonctions  $y(x) = Ce^{-x^2} + \frac{1}{2}$  pour  $C \in \mathbb{R}$ .

### Exercice 4.26

Dans un circuit électrique de type résistance-inductance, le courant  $I$  évolue avec le temps selon

$$I'(t) + \frac{R}{L}I(t) = \frac{V}{L}$$

où  $R$ ,  $L$  et  $V$  sont des constantes associées aux composantes électriques. Résolvez l'équation différentielle. La solution  $I$  tend-elle vers une limite finie ?

#### Correction

On a une EDO linéaire d'ordre 1 avec  $a(t) = 1$ ,  $b(t) = \frac{R}{L}$  et  $g(t) = \frac{V}{L}$ . Donc

- \*  $A(t) = \int \frac{R}{L} \, dt = \frac{R}{L}t,$
- \*  $B(t) = \int \frac{V}{L} e^{A(t)} \, dt = \frac{V}{L} \int e^{\frac{R}{L}t} \, dt = \frac{V}{L} \frac{L}{R} \int \frac{R}{L} e^{-\frac{R}{L}t} \, dt = \frac{V}{R} e^{\frac{R}{L}t},$

donc toutes les solutions de l'EDO sont les fonctions  $I(t) = \alpha e^{-A(t)} + B(t)e^{-A(t)} = \alpha e^{-\frac{R}{L}t} + \frac{V}{R}$  pour  $\alpha \in \mathbb{R}$  et  $I(t) \xrightarrow[t \rightarrow +\infty]{} \frac{V}{R}$ .

### Exercice 4.27 («Les experts - Toulon»)

Le corps de la victime a été trouvé sur le lieu du crime à 2H20 de nuit. Après une demi-heure la température du corps est de 15°C. Quand a eu lieu l'homicide si à l'heure de la découverte la température du corps est de 20°C et si la température externe est de -5°C ?

**Correction**

La loi de Newton affirme qu'il existe une constante  $\gamma < 0$  telle que la température du corps suit l'EDO

$$T'(t) = \gamma(T(t) - T_{\text{ext}}).$$

On commence par calculer toutes les solutions de l'EDO. Étant une équation différentielle du premier ordre, la famille de solutions dépendra d'une constante  $D$  qu'on fixera en utilisant la CI.

Si on réécrit l'EDO sous la forme  $T'(t) - \gamma T(t) = -\gamma T_{\text{ext}}$ , on a une EDO linéaire d'ordre 1 avec  $a(t) = 1$ ,  $b(t) = -\gamma$  et  $g(t) = -\gamma T_{\text{ext}}$ . Donc

- \*  $A(t) = \int -\gamma dt = -\gamma t,$
- \*  $B(t) = \int -\gamma T_{\text{ext}} e^{A(t)} dt = T_{\text{ext}} \int -\gamma e^{-\gamma t} dt = T_{\text{ext}} e^{-\gamma t},$

donc toutes les solutions de l'EDO sont les fonctions  $T(t) = D e^{\gamma t} + T_{\text{ext}}$  pour  $D \in \mathbb{R}$ .

La valeur numérique de la constante d'intégration  $D$  est obtenue grâce à la CI :  $T_0 = T(0) = T_{\text{ext}} + D e^{\gamma \cdot 0}$  donc  $D = T_0 - T_{\text{ext}}$ . Ici  $T_{\text{ext}} = -5^\circ\text{C}$  et  $T_0 = 20^\circ\text{C}$  donc la température du cadavre suit la loi

$$T(t) = -5 + 25 e^{\gamma t}.$$

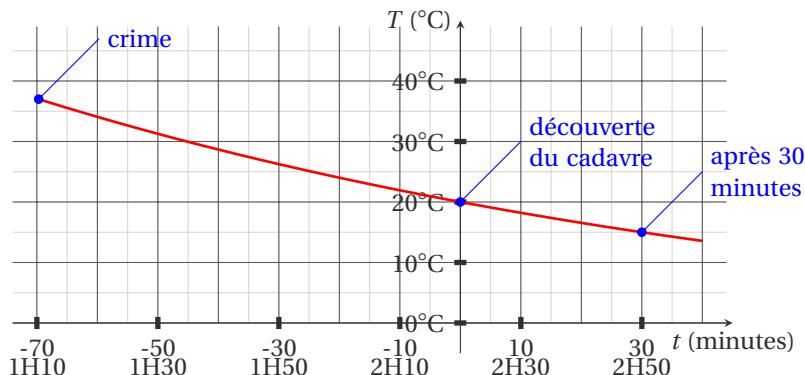
On sait que  $15 = T(30) = -5 + 25 e^{30\gamma}$  d'où  $\gamma = \frac{\ln(4/5)}{30}$ . La température du corps suit donc la loi

$$T(t) = -5 + 25 e^{\frac{\ln(4/5)}{30} t}.$$

Pour déterminer l'heure du meurtre il faut donc résoudre l'équation

$$37 = -5 + 25 e^{\frac{\ln(4/5)}{30} t}$$

d'où  $t = 30 \frac{\ln(42/25)}{\ln(4/5)} \sim -69,7$  minutes, c'est-à-dire à 1H10 de la nuit.

**Exercice 4.28 («Un gâteau presque parfait»)**

Un gâteau est sorti du four à 17H00 quand il est brûlant ( $100^\circ\text{C}$ ). Après 10 minutes sa température est de  $80^\circ\text{C}$  et de  $65^\circ\text{C}$  à 17H20. Déterminer la température de la cuisine.

**Correction**

La loi de Newton affirme qu'il existe une constante  $\gamma < 0$  telle que la température du gâteau suit l'EDO

$$T'(t) = \gamma(T(t) - T_{\text{ext}}).$$

On commence par calculer toutes les solutions de l'EDO. Étant une équation différentielle du premier ordre, la famille de solutions dépendra d'une constante  $D$  qu'on fixera en utilisant la CI.

Si on réécrit l'EDO sous la forme  $T'(t) - \gamma T(t) = -\gamma T_{\text{ext}}$ , on a une EDO linéaire d'ordre 1 avec  $a(t) = 1$ ,  $b(t) = -\gamma$  et  $g(t) = -\gamma T_{\text{ext}}$ . On pose

- \*  $A(t) = \int -\gamma dt = -\gamma t,$
- \*  $B(t) = \int -\gamma T_{\text{ext}} e^{A(t)} dt = T_{\text{ext}} \int -\gamma e^{-\gamma t} dt = T_{\text{ext}} e^{-\gamma t},$

donc toutes les solutions de l'EDO sont les fonctions  $T(t) = D e^{\gamma t} + T_{\text{ext}}$  pour  $D \in \mathbb{R}$ .

La valeur numérique de la constante d'intégration  $D$  est obtenue grâce à la CI :

$$T_0 = T(0) = T_{\text{ext}} + D e^{\gamma \cdot 0} \implies D = T_0 - T_{\text{ext}} \implies T(t) = T_{\text{ext}} + (T_0 - T_{\text{ext}}) e^{\gamma t}.$$

Ici l'inconnue est  $T_{\text{ext}}$ . On sait que  $T(t = 0) = 100^\circ\text{C}$  et  $T(t = 10) = 80^\circ\text{C}$  et  $T(t = 20) = 65^\circ\text{C}$ . Il s'agit donc de résoudre le système de trois équations en les trois inconnues  $\gamma, D, T_{\text{ext}}$ :

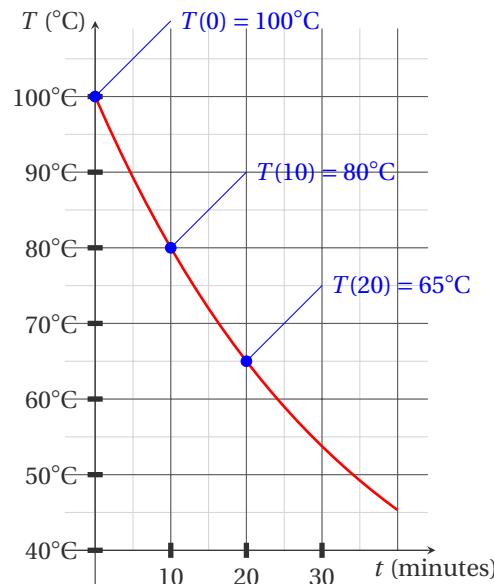
$$\begin{cases} 100 &= T_{\text{ext}} + D, \\ 80 &= T_{\text{ext}} + D e^{10\gamma}, \\ 65 &= T_{\text{ext}} + D e^{20\gamma}. \end{cases}$$

La première équation se réécrit  $D = 100 - T_{\text{ext}}$ , la seconde équation se réécrit  $e^{10K} = \frac{80 - T_{\text{ext}}}{D} = \frac{80 - T_{\text{ext}}}{100 - T_{\text{ext}}}$ , la troisième  $e^{20K} = \frac{65 - T_{\text{ext}}}{D} = \frac{65 - T_{\text{ext}}}{100 - T_{\text{ext}}}$  donc

$$\frac{80 - T_{\text{ext}}}{100 - T_{\text{ext}}} = e^{10K} = \frac{e^{20K}}{e^{10K}} = \frac{65 - T_{\text{ext}}}{80 - T_{\text{ext}}}$$

d'où  $(80 - T_{\text{ext}})^2 = (65 - T_{\text{ext}})(100 - T_{\text{ext}})$ , ainsi  $T_{\text{ext}} = \frac{65 \times 100 - 80^2}{5} = 20$ . La cuisine est donc à  $20^\circ\text{C}$  et, plus généralement, la température du gâteau évolue selon la loi

$$T(t) = 20 + 80e^{\frac{\ln(3/4)}{10}t}.$$



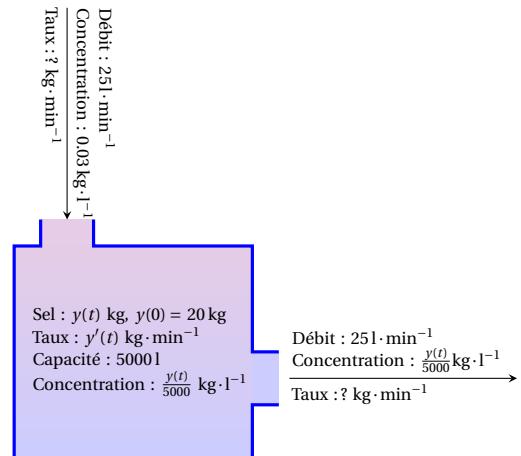
### Exercice 4.29

On considère un réservoir de capacité 5000 l rempli d'une solution sel/eau parfaitement mélangée contenant 20 kg de sel. Un mélange qui contient 0.03 kg de sel par litre d'eau entre dans le réservoir à un débit de  $251 \cdot \text{min}^{-1}$ . La solution est maintenue bien mélangée. Si  $y(t)$  désigne la quantité (en kilos) de sel dissoute dans le réservoir à l'instant  $t$ ,  $y'(t)$  représente le taux de variation de la quantité de sel, i.e. la différence entre le taux auquel le sel entre et le taux auquel il en sort.

- Après avoir calculé les taux auxquels le sel entre et sort du réservoir, montrer que cette situation est décrite par le problème de Cauchy

$$\begin{cases} y'(t) = 0.75 - \frac{y(t)}{200}, \\ y(0) = 20. \end{cases}$$

- Calculer l'unique solutions de ce problème.
- Combien de sel reste dans le réservoir après une demi-heure ?



### Correction

- Le taux auquel le sel entre est  $(0.03 \text{ kg})(251 \cdot \text{min}^{-1}) = 0.75 \text{ kg} \cdot \text{min}^{-1}$ . Comme le réservoir contient constamment 5000 l de liquide, la concentration est égale à  $y(t)/5000$  (exprimée en  $\text{kg} \cdot \text{l}^{-1}$ ). Le débit du mélange qui sort est alors de  $251 \cdot \text{min}^{-1}$ , donc le taux auquel le sel sort est  $(\frac{y(t)}{5000} \text{ kg} \cdot \text{l}^{-1})(251 \cdot \text{min}^{-1}) = \frac{y(t)}{200} \text{ kg} \cdot \text{min}^{-1}$ . L'équation différentielle qui décrit cette variation s'écrit alors

$$y'(t) = 0.75 - \frac{y(t)}{200}$$

- On a une EDO linéaire d'ordre 1 avec  $a(t) = 1$ ,  $b(t) = 1/200$ ,  $g(t) = 0.75$ . On pose

$$\begin{aligned} * \quad A(t) &= \int \frac{b(t)}{a(t)} dt = \int \frac{1}{200} dt = \frac{1}{200} t, \\ * \quad B(t) &= \int \frac{g(t)}{a(t)} e^{A(t)} dt = 0.75 \int e^{t/200} dt = 150 e^{t/200}, \end{aligned}$$

donc toutes les solutions de l'EDO sont les fonctions  $y(t) = Ee^{-t/200} + 150$  pour  $E \in \mathbb{R}$ .

La valeur numérique de la constante d'intégration  $E$  est obtenue grâce à la CI :  $20 = y(0) = E + 150$  donc  $E = -130$  et l'unique solution du problème de CAUCHY est

$$y(t) = 150 - 130e^{-t/200}.$$

- Reste à calculer la quantité de sel après 30 minutes :  $y(30) = 150 - 130e^{-3/20} \simeq 38.1 \text{ kg}$ .

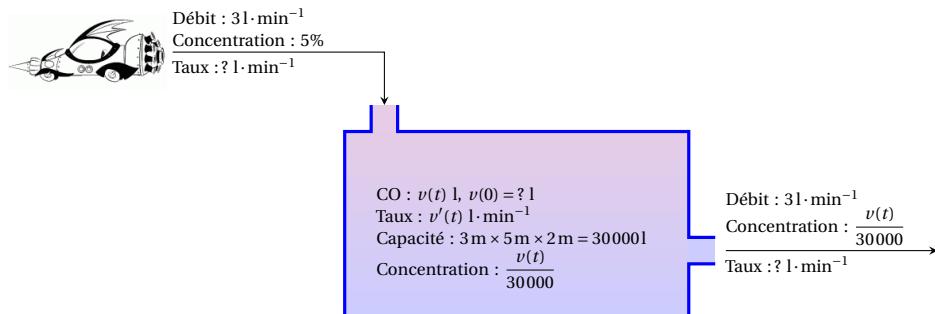
### Exercice 4.30

L'air d'un garage de  $3\text{ m} \times 5\text{ m} \times 2\text{ m}$  est initialement chargée de 0.001% de monoxyde de carbone (CO). À l'instant  $t = 0$ , on fait tourner un moteur et des fumées toxiques contenant 5% de CO se dégagent de la pièce à raison de 3 litres par minute. Heureusement, l'air de la pièce est éliminée à la même vitesse de  $31 \cdot \text{min}^{-1}$ . On note  $v(t)$  le volume de CO présent dans la pièce au temps  $t$ .

- En supposant que le mélange se fait instantanément, montrer que cette situation est décrite par le problème de Cauchy

$$\begin{cases} v'(t) = 0.15 - \frac{v(t)}{10000}, \\ v(0) = 0.3. \end{cases}$$

- Déterminer le volume  $v(t)$  de CO présent dans la pièce au temps  $t$ . Calculer vers quelle valeur limite  $v(t)$  tend lorsque  $t$  tend vers l'infini.
- Le seuil critique pour la santé est de 0.015% de CO. Après combien de temps ce taux est-il atteint ?



#### Correction

- Le taux de CO produit par minute est  $0.05 \times 31 \cdot \text{min}^{-1} = 0.151 \cdot \text{min}^{-1}$ . Le débit de l'air qui sort est de  $31 \cdot \text{min}^{-1}$ , donc le taux auquel le CO sort est  $\frac{v(t)}{30000} \times 31 \cdot \text{min}^{-1} = \frac{v(t)}{10000} \cdot \text{min}^{-1}$ . L'équation différentielle qui décrit cette variation s'écrit alors

$$v'(t) = 0.15 - \frac{v(t)}{10000}.$$

À l'instant  $t = 0$  le volume de CO présent dans le garage est  $0.001\% \times 30000\text{l} = 0.3\text{l}$ .

- On a une EDO linéaire d'ordre 1 avec  $a(t) = 1$ ,  $b(t) = 1/10000$ ,  $g(t) = 0.15$ . On pose

$$\begin{aligned} * \quad A(t) &= \int \frac{b(t)}{a(t)} dt = \int \frac{1}{10000} dt = \frac{1}{10000} t, \\ * \quad B(t) &= \int \frac{g(t)}{a(t)} e^{A(t)} dt = 0.15 \int e^{t/10000} dt = 1500 e^{t/10000}, \end{aligned}$$

donc toutes les solutions de l'EDO sont les fonctions  $v(t) = Ee^{-t/10000} + 1500$  pour  $E \in \mathbb{R}$ .

La valeur numérique de la constante d'intégration  $E$  est obtenue grâce à la CI :  $0.3 = v(0) = E + 1500$  donc  $E = -(1500 - 0.3)$ . L'unique solution du problème de CAUCHY est donc

$$v(t) = 1500 - (1500 - 0.3)e^{-t/10000} \xrightarrow[t \rightarrow +\infty]{} 1500.$$

- Reste à calculer après combien de minutes le taux de CO atteint 0.015% :  $0.00015 = 1500 - (1500 - 0.3)e^{-t/10000}$  ssi  $t = 10000 \ln\left(\frac{1500 - 0.3}{1500 - 4.5}\right) \simeq 28.04 \text{ min}$ .

### Exercice 4.31 (Un escargot sur un élastique)

Un escargot avance d'un mètre par jour sur un élastique d'un kilomètre de long. Mais l'élastique s'étire d'un kilomètre par jour. L'escargot arrivera-t-il au bout de l'élastique ?

Source : <http://allken-bernard.org/pierre/weblog/?p=209>

#### Correction

On note  $L(t)$  la longueur de l'élastique à l'instant  $t$  et  $\ell(t)$  la distance parcourue par l'escargot à l'instant  $t$ . Pour les unités de mesure, on convient qu'une unité de temps correspond à un jour et les longueurs sont mesurées en mètres. On a  $L(0) = 1000$ ,  $\ell(0) = 0$  et il s'agit de voir si  $\ell(t) = L(t)$  pour un certain  $t$ .

Pour tout  $t \geq 0$ ,

$$L(t) = 1000t + 1000$$

et on peut définir  $y(t)$  la fraction de l'élastique parcourue par l'escargot à l'instant  $t$  :

$$y(t) = \frac{\ell(t)}{L(t)} \quad \forall t \geq 0.$$

La vitesse  $\ell'$  de l'escargot par rapport à l'extrémité fixe de l'élastique est la somme de deux vitesses : la vitesse de l'escargot sur l'élastique, soit 1 mètre par jour, et la vitesse du point de l'élastique où se trouve l'escargot (on peut faire l'hypothèse que cette vitesse est proportionnelle à l'abscisse de l'escargot) :

$$\ell'(t) = 1 + y(t)L'(t) \quad \forall t \geq 0,$$

donc

$$y'(t) = \frac{\ell'(t)}{L(t)} - y(t)\frac{L'(t)}{L(t)} = \frac{1 + y(t)L'(t)}{L(t)} - y(t)\frac{L'(t)}{L(t)} = \frac{1}{L(t)} \quad \forall t \geq 0.$$

Puisque  $y(0) = 0$ , on en conclut que

$$y(t) = \int_0^t \frac{1}{L(\tau)} d\tau = \frac{1}{1000} \int_0^t \frac{1}{1+\tau} d\tau = \frac{1}{1000} [\ln(1+\tau)]_0^t = \frac{1}{1000} \ln(1+t)$$

et donc

$$\ell(t) = y(t)L(t) = \frac{1000t+1000}{1000} \ln(1+t) = (1+t)\ln(1+t) \quad \forall t \geq 0.$$

L'escargot touchera l'extrémité mobile de l'élastique lorsque  $\ell(t) = L(t)$ , c'est-à-dire à l'instant  $t_f = e^{1000} - 1 \approx 1.97 \times 10^{434}$  jours  $\approx 5.397 \times 10^{431}$  années (ce qui correspond à  $\approx 3.9 \times 10^{421}$  fois l'âge de l'univers).

En étudiant la fonction  $t \mapsto d(t) = L(t) - \ell(t)$ , on trouve que cette distance est maximale<sup>4</sup> à l'instant  $t_0 = e^{999} - 1$ ; après cet instant l'escargot commence à se rapprocher de l'extrémité de l'élastique pour en arriver au but à l'instant  $t_f = e^{1000} - 1$ . À l'instant  $t_0$  l'escargot se déplace à une vitesse de 1000 kilomètre par jour et a parcouru  $y(t_0) = 99.9\%$  de l'élastique, elle a parcouru 99.9% de l'élastique mais elle n'a jamais été aussi loin de son but!

## Calcul analytique des solutions d'une EDO de type Bernoulli

### Exercice 4.32

Déterminer la solution générale des EDO suivantes après avoir indiqué sur quelle intervalle la solution est définie :

a)  $y'(t) - \frac{1}{t}y(t) = (y(t))^3 \sin(t)$       b)  $y'(t) + ty(t) = t^3(y(t))^2$

#### Correction

(a) L'EDO  $y'(t) - \frac{1}{t}y(t) = (y(t))^3 \sin(t)$  est une équation différentielle de BERNOULLI. Comme  $u(t) = 1$  pour tout  $t \in \mathbb{R}^*$ , on cherche sa solution générale sur  $]-\infty; 0[$  et sur  $]0; +\infty[$ .

$$\star A(t) = (1-\alpha) \int \frac{v(t)}{u(t)} dt = \int \frac{1}{t} dt = 2 \ln|t|,$$

$$\star B(t) = (1-\alpha) \int \frac{w(t)}{u(t)} e^{A(t)} dt = - \int t^2 \sin(t) dt = 2t^2 \cos(t) - 4t \sin(t) - 4 \cos(t),$$

$$\star z(t) = (C_{1,2} + B(t)) e^{-A(t)} = (C_{1,2} + 2t^2 \cos(t) - 4t \sin(t) - 4 \cos(t)) e^{-2 \ln|t|} = \frac{C_{1,2} + 2t^2 \cos(t) - 4t \sin(t) - 4 \cos(t)}{t^2},$$

$$\star y(t) = (z(t))^{-1/2} = \frac{1}{\sqrt{z(t)}}$$

et on conclut que la solution générale de l'EDO de BERNOULLI assignée est

$$y: \mathbb{R}^* \rightarrow \mathbb{R}$$

$$t \mapsto \begin{cases} \frac{t}{\sqrt{C_1 + 2t^2 \cos(t) - 4t \sin(t) - 4 \cos(t)}} & \text{si } t < 0 \quad \text{avec } C_1 \in \mathbb{R}^+, \\ \frac{-t}{\sqrt{C_2 + 2t^2 \cos(t) - 4t \sin(t) - 4 \cos(t)}} & \text{si } t > 0 \quad \text{avec } C_2 \in \mathbb{R}^+, \end{cases}$$

(b) L'EDO  $y'(t) + ty(t) = t^3(y(t))^2$  est une équation différentielle de BERNOULLI. Comme  $u(t) = 1$  pour tout  $t \in \mathbb{R}$ , on cherche sa solution générale sur  $\mathbb{R}$ .

$\star$  *Solution nulle* : la fonction  $y(t) = 0$  pour tout  $t \in \mathbb{R}$  est solution de l'EDO donnée. Toute autre solution ne s'annule jamais. Supposons dans la suite que  $y(t) \neq 0$  pour tout  $t \in \mathbb{R}$ .

4.  $d'(t) = 999 - \ln(1+t)$

- \*  $A(t) = (1 - \alpha) \int \frac{v(t)}{u(t)} dt = - \int t dt = -\frac{t^2}{2},$
- \*  $B(t) = (1 - \alpha) \int \frac{w(t)}{u(t)} e^{A(t)} dt = - \int t^3 e^{-t^2/2} dt = -2 \int xe^x dx = -2(x-1)e^x = (2+t^2)e^{-t^2/2},$
- \*  $z(t) = ce^{t^2/2} + 2 + t^2,$
- \*  $y(t) = (z(t))^{-1} = \frac{1}{z(t)}$

et on conclut que la solution générale de l'EDO de BERNOULLI assignée est

$$y: \mathbb{R} \rightarrow \mathbb{R}$$

$$t \mapsto \frac{1}{Ce^{t^2/2} + t^2 + 2} \quad \text{avec } C \in \mathbb{R}$$

## Approximation numérique d'EDO

### ★ Exercice 4.33

Considérons le problème de CAUCHY

trouver une fonction  $y: I \subset \mathbb{R} \rightarrow \mathbb{R}$  définie sur un intervalle  $I = [t_0, T]$  telle que

$$\begin{cases} y'(t) = \varphi(t, y(t)), & \forall t \in I = [t_0, T], \\ y(t_0) = y_0, \end{cases}$$

avec  $y_0$  une valeur donnée et supposons que l'on ait montré l'existence et l'unicité d'une solution  $y$  pour  $t \in I$ .

Pour  $h = (T - t_0)/N > 0$  soit  $t_n \equiv t_0 + nh$  avec  $n = 0, 1, 2, \dots, N$  une suite de noeuds de  $I$  induisant une discréttisation de  $I$  en sous-intervalles  $I_n = [t_n; t_{n+1}]$ . La longueur  $h$  est appelé le *pas de discréttisation*.

Pour chaque noeud  $t_n$ , on cherche la valeur inconnue  $u_n$  qui approche la valeur exacte  $y(t_n)$ . L'ensemble des valeurs  $\{u_0 = y_0, u_1, \dots, u_{N_h}\}$  représente la solution numérique.

Une méthode classique, la **méthode d'EULER explicite** (ou **progressive**, de l'anglais *forward*), consiste à construire une solution numérique ainsi

$$\begin{cases} u_0 = y(t_0) = y_0, \\ u_{n+1} = u_n + h\varphi(t_n, u_n), & n = 0, 1, 2, \dots, N_h - 1. \end{cases}$$

Une autre méthode classique, la **méthode d'EULER modifiée**, consiste à construire une solution numérique ainsi

$$\begin{cases} u_0 = y(t_0) = y_0, \\ u_{n+1} = u_n + h\varphi(t_n + \frac{h}{2}, u_n + \frac{h}{2}\varphi(t_n, u_n)), & n = 0, 1, 2, \dots, N_h - 1. \end{cases}$$

- ① On se propose d'écrire une **function** pour résoudre numériquement un problème de CAUCHY avec la méthode d'EULER explicite. La **function** prend en entrée **t0** et **T** les extrêmes de l'intervalle d'intégration, **y0** la donnée initiale, **N** le nombre de sous-intervalles qu'on va considérer et **phi** une chaîne contenant l'expression de  $\varphi(t, y)$  et elle donne en sortie **u** le vecteur contenant l'approximation de  $y$  en chaque point  $t_n$ .
- ② À l'aide de la fonction ainsi écrite résoudre numériquement l'équation différentielle  $y'(t) = y(t)$  sur l'intervalle  $[0; 2]$  en supposant  $y(0) = 1$  pour différentes valeurs du paramètre  $N$ : 10, 50, 100.
- ③ Reprendre les deux questions précédentes avec le schéma d'EULER modifié.
- ④ Tracer, avec les deux schémas ci-dessus, la solution de l'équation différentielle  $y'(t) = \frac{y(t)}{2(t+1)}$  et  $y(0) = 1$  sur l'intervalle  $[0; 8]$ .

### Correction

- ① La **function** **eulerexplicite** prend en entrée **t0** et **T** les extrêmes de l'intervalle d'intégration, **y0** la donnée initiale, **N** le nombre de sous-intervalles qu'on va considérer et **phi** une chaîne contenant l'expression de  $\varphi(t, y)$  et elle donne en sortie **u** le vecteur contenant l'approximation de  $y$  en chaque point  $t_n$ .

```
function [u]=eulerexplicite(t0,T,y0,N,phi)
t=zeros(N,1);
u=zeros(N,1);
t(1)=t0;
u(1)=y0;
h=(T-t0)/N;
t=linspace(t0,T,N);
```

```

for n=1:N-1
    phin=feval(phi,t(n),u(n));
    u(n+1) = u(n)+h*phin;
end
end

```

```

② t0=0;
T=2;
y0=1;
phi=@(t,y)[y];

N=10;
t10=linspace(t0,T,N)
u10=eulerexplicite(t0,T,y0,N,phi);

N=50;
t50=linspace(t0,T,N)
u50=eulerexplicite(t0,T,y0,N,phi);

N=100;
t100=linspace(t0,T,N)
u100=eulerexplicite(t0,T,y0,N,phi);

plot(t10,u10,t50,u50,t100,u100)

```

- ③ La `function eulermodifie` prend en entrée  $t_0$  et  $T$  les extrêmes de l'intervalle d'intégration,  $y_0$  la donnée initiale,  $N$  le nombre de sous-intervalles qu'on va considérer et  $\phi$  une chaîne contenant l'expression de  $\varphi(t, y)$  et elle donne en sortie  $u$  le vecteur contenant l'approximation de  $y$  en chaque point  $t_n$ .

```

function [u]=eulermodifie(t0,T,y0,N,phi)
t=zeros(N,1);
u=zeros(N,1);
t(1)=t0;
u(1)=y0;
h=(T-t0)/N;
t=linspace(t0,T,N);
for n=1:N-1
    phin=feval(phi,t(n),u(n));
    utemp=u(n)+h/2*phin;
    phitemp=feval(phi,t(n)+h/2,utemp);
    u(n+1) = u(n)+h*phitemp;
end
end

```

```

④ t0=0;
T=2;
y0=1;
phi=@(t,y)[y];

N=10;
t10=linspace(t0,T,N)
u10=eulermodifie(t0,T,y0,N,phi);

N=50;
t50=linspace(t0,T,N)
u50=eulermodifie(t0,T,y0,N,phi);

N=100;
t100=linspace(t0,T,N)
u100=eulermodifie(t0,T,y0,N,phi);

plot(t10,u10,t50,u50,t100,u100)

```

### ★ Exercice 4.34

Considérons le problème de CAUCHY

trouver la fonction  $y: I \subset \mathbb{R} \rightarrow \mathbb{R}$  définie sur l'intervalle  $I = [0, 1]$  telle que

$$\begin{cases} y'(t) = y(t), & \forall t \in I = [0, 1], \\ y(0) = 1 \end{cases}$$

dont la solution est  $y(t) = e^t$ . On le résout avec la méthode d'EULER explicite avec différentes valeurs de  $N$ , à savoir  $2, 2^2, 2^3, \dots, 2^{12}$  (ce qui correspond à différentes valeurs de  $h$ , à savoir  $1/2, 1/4, 1/8, \dots, 1/4096$ ). Pour chaque valeur de  $N$ , on ne sauvegarde que l'erreur commise au point final  $t = 1$  et on stocke tous ces erreurs dans le vecteur  $\text{err}$  de sorte que  $\text{err}[k]$  contient  $|y(t=1) - u_{N(k)}|$  avec  $N(k) = 2^{k+1}$ .

Pour estimer l'ordre de convergence  $p$  on calculera la pente de la droite de régression sur l'ensemble de points  $\{\ln(h_k), \ln(\text{err}_k)\}_{k=1}^N$ . En effet, si l'erreur  $\text{err}$  est égale à  $Ch^p$  alors  $\ln(\text{err}) = \ln(C) + p \ln(h)$ : en échelle logarithmique,  $p$  représente donc la pente de la ligne droite  $\ln(\text{err})$ .

### Correction

On initialise les données :

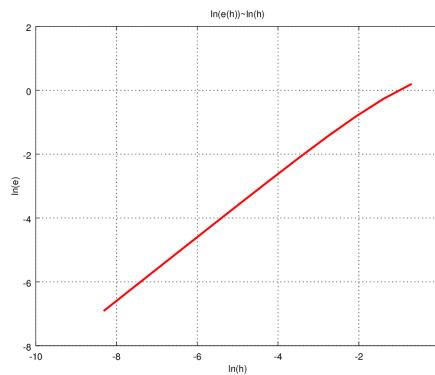
```
t0=0;
T=1;
y0=1;
phi=@(t,y) [y];
exacte=@(t) [exp(t)];
```

Pour  $N = 2^k$ ,  $k = 1, \dots, 12$ , on calcule la solution approchée par la méthode d'Euler explicite et on évalue l'erreur en  $t = 1$ :

```
for k=1:12
    u=eulerexplicite(t0,T,y0,2^k,phi); % approximation de y(t) pour t=[t0,t0+h,...,T]
    uFIN=u(end) % approximation de y(t=T)
    errFIN(K)=abs(exacte(T)-uFIN);
end
```

Pour afficher l'ordre de convergence  $p$  on utilise une échelle logarithmique, *i.e.* on représente  $\ln(h)$  sur l'axe des abscisses et  $\ln(e)$  sur l'axe des ordonnées. Le but de cette représentation est clair : si  $e = Ch^p$  alors  $\ln(e) = \ln(C) + p \ln(h)$ . En échelle logarithmique,  $p$  représente donc la pente de la ligne droite  $\ln(e)$ :

```
h=2.^(-[1:Kmax]);
plot(log(h),log(errFIN),'LineWidth',2,'r-') % equivalent a loglog(h,errFIN,'LineWidth',2,'r-')
polyfit(log(h),log(errFIN),1)(1) % premier coefficient = pente de la droite
title('ln(err(h))~ln(h)')
xlabel('ln(h)')
ylabel('ln(err)')
grid
```



### ★ Exercice 4.35

L'évolution de la concentration de certaines réactions chimiques au cours du temps peut être décrite par l'équation différentielle

$$y'(t) = -\frac{1}{1+t^2} y(t).$$

Sachant qu'à l'instant  $t = 0$  la concentration est  $y(0) = 5$ , déterminer la concentration à  $t = 2$  à l'aide de la méthode d'EULER implicite avec un pas  $h = 0.5$ .

**Correction**

La méthode d'EULER implicite est une méthode d'intégration numérique d'EDO du premier ordre de la forme  $y'(t) = F(t, y(t))$ . C'est une méthode itérative : en choisissant un pas de discréttisation  $h$ , la valeur  $y$  à l'instant  $t + h$  se déduit de la valeur de  $y$  à l'instant  $t$  par l'approximation linéaire

$$y(t+h) \approx y(t) + hy'(t+h) = y(t) + hF(t+h, y(t+h)).$$

On pose alors  $t_n = t_0 + nh$ ,  $n \in \mathbb{N}$ . En résolvant l'équation non-linéaire

$$u_{n+1} = u_n + hF(t_{n+1}, u_{n+1}),$$

on obtient une suite  $(u_n)_{n \in \mathbb{N}}$  qui approche les valeurs de la fonction  $y$  en  $t_n$ . Dans notre cas, l'équation non-linéaire s'écrit

$$u_{n+1} = u_n - \frac{h}{1+t_{n+1}^2} u_{n+1}.$$

Elle peut être résolue algébriquement et cela donne la suite

$$u_{n+1} = \frac{u_n}{1 + \frac{h}{1+t_{n+1}^2}}.$$

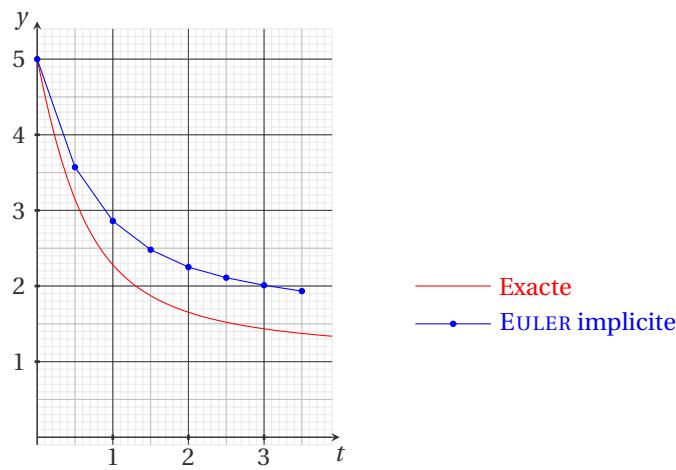
Si à l'instant  $t = 0$  la concentration est  $y(0) = 5$ , et si  $h = 1/2$ , alors  $t_n = n/2$  et

$$u_{n+1} = \frac{4 + (n+1)^2}{6 + (n+1)^2} u_n.$$

On obtient donc

$n$	$t_n$	$u_n$
0	0	5
1	0.5	$\frac{4+1^2}{6+1^2} 5 = \frac{5}{7} 5 = \frac{25}{7} \approx 3.57$
2	1.0	$\frac{4+2^2}{6+2^2} \frac{25}{7} = \frac{8}{10} \frac{25}{7} = \frac{20}{7} \approx 2.86$
3	1.5	$\frac{4+3^2}{6+3^2} \frac{20}{7} = \frac{13}{15} \frac{20}{7} = \frac{52}{21} \approx 2.48$
4	2.0	$\frac{4+4^2}{6+4^2} \frac{52}{21} = \frac{20}{22} \frac{52}{21} = \frac{520}{231} \approx 2.25$

La concentration à  $t = 2$  est d'environ 2.25 qu'on peut comparer avec le calcul exact  $y(2) = 5e^{-\arctan(2)} \approx 1.652499838$ .

**★ Exercice 4.36 (Loi de NEWTON ☕)**

Considérons une tasse de café à la température de  $75^\circ\text{C}$  dans une salle à  $25^\circ\text{C}$ . On suppose que la température du café suit la loi de Newton, c'est-à-dire que la vitesse de refroidissement du café est proportionnelle à la différence des températures. En formule cela signifie qu'il existe une constante  $K < 0$  telle que la température vérifie l'équation différentielle ordinaire (EDO) du premier ordre.

$$T'(t) = K(T(t) - 25).$$

La condition initiale (CI) est donc simplement

$$T(0) = 75.$$

Pour calculer la température à chaque instant on a besoin de connaître la constante  $K$ . Cette valeur peut être déduite en constatant qu'après 5 minutes le café est à 50°C, c'est-à-dire

$$T(5) = 50.$$

On peut montrer que la température du café évolue selon la fonction

$$T(t) = 25 + 50e^{-\frac{\ln(2)}{5}t}.$$

Comparer cette solution avec la solution approchée obtenue par la méthode d'EULER explicite.

### Correction

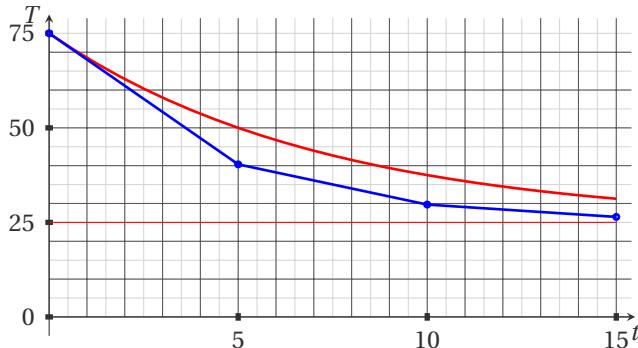
Supposons de connaître  $K$  mais de ne pas vouloir/pouvoir calculer la fonction  $T(t)$ . Grâce à la méthode d'EULER on peut estimer la température à différentes instants  $t_i$  en faisant une discrétisation temporelle du futur (i.e. on construit une suite de valeurs  $\{t_i = 0 + i\Delta t\}_i$ ) et en construisant une suite de valeurs  $\{T_i\}_i$  où chaque  $T_i$  est une approximation de  $T(t_i)$ . Si on utilise la méthode d'EULER, cette suite de température est ainsi construite :

$$\begin{cases} T_{i+1} = T_i - \frac{\ln(2)}{5}\Delta t (T_i - 25), \\ T_0 = 75, \end{cases}$$

qu'on peut réécrire comme

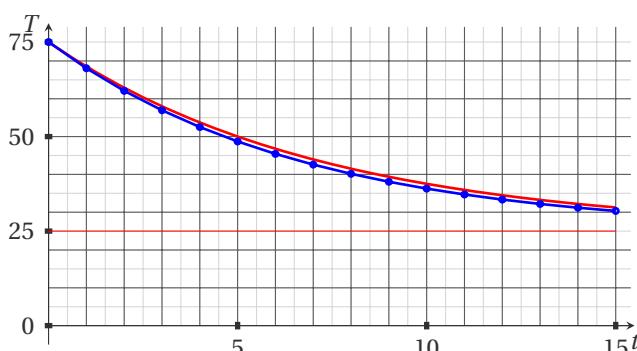
$$\begin{cases} T_{i+1} = (1 - \frac{\ln(2)}{5}\Delta t)T_i + 5\ln(2)\Delta t, \\ T_0 = 75. \end{cases}$$

1. Exemple avec  $\Delta t = 5$  :



$t_i$	$T(t_i)$	$T_i$	$T(t_i) - T_i$
0.000000	75.000000	75.000000	0.000000
5.000000	50.000000	40.342641	9.657359
10.000000	37.500000	29.707933	7.792067
15.000000	31.250000	26.444642	4.805358

2. Exemple avec  $\Delta t = 1$  :



$t_i$	$T(t_i)$	$T_i$	$T(t_i) - T_i$
0.000000	75.000000	75.000000	0.000000
1.000000	68.527528	68.068528	0.459000
2.000000	62.892914	62.097962	0.794952
3.000000	57.987698	56.955093	1.032605
4.000000	53.717459	52.525176	1.192283
5.000000	50.000000	48.709377	1.290623
6.000000	46.763764	45.422559	1.341205
7.000000	43.946457	42.591391	1.355066
8.000000	41.493849	40.152707	1.341142
9.000000	39.358729	38.052095	1.306634
10.000000	37.500000	36.242691	1.257309
11.000000	35.881882	34.684123	1.197759
12.000000	34.473229	33.341618	1.131610
13.000000	33.246924	32.185225	1.061700
14.000000	32.179365	31.189141	0.990224
15.000000	31.250000	30.331144	0.918856

### ★ Exercice 4.37 («Les experts - Toulon»)

La loi de Newton affirme que la vitesse de refroidissement d'un corps est proportionnelle à la différence entre la température du corps et la température externe, autrement dit qu'il existe une constante  $K < 0$  telle que la température du corps suit l'équation différentielle

$$\begin{cases} T'(t) = K(T(t) - T_{\text{ext}}), \\ T(0) = T_0. \end{cases}$$

1. Soit  $\Delta t$  le pas temporel. Écrire le schéma d'EULER implicite pour approcher la solution de cette équation différentielle.
2. Soit  $T_{\text{ext}} = 0^\circ\text{C}$ . En déduire une forme du type

$$T_{n+1} = g(\Delta t, n, T_0)$$

avec  $g(\Delta t, n, T_0)$  à préciser (autrement dit, l'itéré en  $t_n$  ne dépend que de  $\Delta t$ , de  $n$  et de  $T_0$ ). Que peut-on en déduire sur la convergence de la méthode ?

3. *Problème.* Un homicide a été commis. On veut établir l'heure du crime sachant que

- pour un corps humaine on peut approcher  $K \approx -0.007438118376$  (l'échelle du temps est en minutes et la température en Celsius),
- le corps de la victime a été trouvé sur le lieu du crime à 2H20 du matin,
- à l'heure du décès la température du corps était de  $37^\circ\text{C}$ ,
- à l'heure de la découverte la température du corps est de  $20^\circ\text{C}$ ,
- la température externe est  $T_{\text{ext}} = 0^\circ\text{C}$ .

Approcher l'heure de l'homicide en utilisant le schéma d'EULER implicite avec  $\Delta t = 10$  minutes.

### Correction

1. La méthode d'EULER implicite (ou régressive) est une méthode d'intégration numérique d'EDO du premier ordre de la forme  $T'(t) = F(t, T(t))$ . En choisissant un pas de discréttisation  $\Delta t$ , nous obtenons une suite de valeurs  $(t_n, T_n)$  qui peuvent être une excellente approximation de la fonction  $T(t)$  avec

$$\begin{cases} t_n = t_0 + n\Delta t, \\ T_{n+1} = T_n + F(t_{n+1}, T_{n+1})\Delta t. \end{cases}$$

La méthode d'EULER implicite pour cette EDO s'écrit donc

$$T_{n+1} = T_n + K\Delta t(T_{n+1} - T_{\text{ext}}).$$

2. Si  $T_{\text{ext}} = 0^\circ\text{C}$ , en procédant par récurrence sur  $n$  on obtient

$$T_{n+1} = g(\Delta t, n) = \frac{1}{1 - K\Delta t} T_n = \frac{1}{(1 - K\Delta t)^{n+1}} T_0,$$

autrement dit, l'itérée en  $t_n$  ne dépend que de  $\Delta t$  et de  $n$  mais ne dépend pas de  $T_n$ . Comme  $0 < \frac{1}{1 - K\Delta t} < 1$  pour tout  $\Delta t > 0$ , la suite est positive décroissante ce qui assure que la solution numérique est stable et convergente.

3. On cherche combien de minutes se sont écoulés entre le crime et la découverte du corps, autrement dit on cherche  $n$  tel que

$$20 = \frac{1}{(1 - K\Delta t)^{n+1}} 37 \implies (1 - K\Delta t)^{n+1} = \frac{37}{20} \implies n+1 = \log_{(1-K\Delta t)} \left( \frac{37}{20} \right) = \frac{\ln \left( \frac{37}{20} \right)}{\ln(1 - K\Delta t)} \implies n \approx 8.$$

Comme  $t_n = t_0 + n\Delta t$ , si  $t_n = 2H20$  alors  $t_0 = t_n - n\Delta t = 2H20 - 1H20 = 01H00$ .

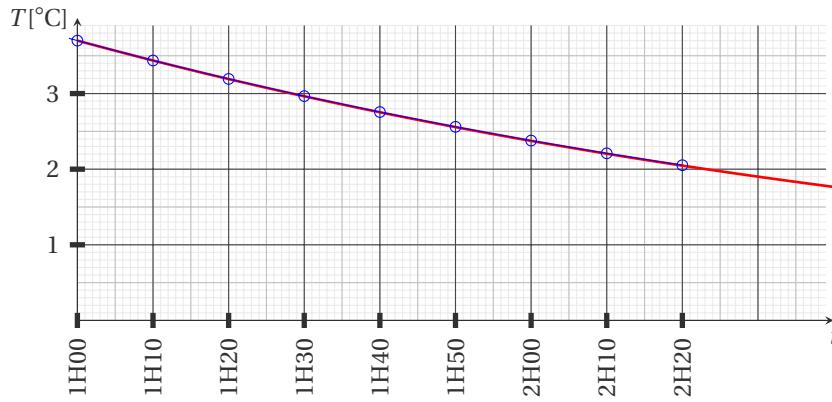
Pour cette équation différentielle, il est possible de calculer analytiquement ses solutions : la température du cadavre suit la loi

$$T(t) = 37e^{Kt}.$$

Pour déterminer l'heure du meurtre il faut alors résoudre l'équation

$$20 = 37e^{Kt}$$

d'où  $t = \frac{1}{K} \ln \frac{20}{37} \approx 82,70715903$  minutes, c'est-à-dire 83 minutes avant 2H20 : le crime a été commis à 00H57.



### Exercice 4.38 (Interpolation, Quadrature et EDO)

1. Soit  $f$  une fonction de classe  $\mathcal{C}^1([-1, 1])$ . Écrire le polynôme  $p \in \mathbb{R}_2[\tau]$  qui interpole  $f$  aux points  $-1, 0$  et  $1$ .
2. Construire une méthode de quadrature comme suit :

$$\int_0^1 f(\tau) d\tau \approx \int_0^1 p(\tau) d\tau.$$

NB : on intègre sur  $[0, 1]$  mais on interpole en  $-1, 0$  et  $1$ .

3. À l'aide d'un changement de variable affine entre l'intervalle  $[0, 1]$  et l'intervalle  $[a, b]$ , en déduire une formule de quadrature pour l'intégrale

$$\int_a^b f(x) dx$$

lorsque  $f$  est une fonction de classe  $\mathcal{C}^1([2a - b, b])$ .

Remarque :  $[2a - b, b] = [a - (b - a), a + (b - a)]$

4. Considérons le problème de CAUCHY : trouver  $y : [t_0, T] \subset \mathbb{R} \rightarrow \mathbb{R}$  tel que

$$\begin{cases} y'(t) = \varphi(t, y(t)), & \forall t \in [t_0, T], \\ y(t_0) = y_0, \end{cases}$$

dont on suppose l'existence d'une unique solution  $y$ .

On subdivise l'intervalle  $[t_0; T]$  en  $N$  intervalles  $[t_n; t_{n+1}]$  de largeur  $h = \frac{T - t_0}{N}$  avec  $t_n = t_0 + nh$  pour  $n = 0, \dots, N$ . Utiliser la formule obtenue au point 3 pour approcher l'intégrale

$$\int_{t_n}^{t_{n+1}} \varphi(t, y(t)) dt.$$

En déduire un schéma à deux pas implicite pour l'approximation de la solution du problème de CAUCHY.

#### Correction

1. On cherche les coefficients  $\alpha, \beta$  et  $\gamma$  du polynôme  $p(\tau) = \alpha + \beta\tau + \gamma\tau^2$  tels que

$$\begin{cases} p(-1) = f(-1), \\ p(0) = f(0), \\ p(1) = f(1), \end{cases} \quad \text{c'est à dire} \quad \begin{cases} \alpha - \beta + \gamma = f(-1), \\ \alpha = f(0), \\ \alpha + \beta + \gamma = f(1). \end{cases}$$

Donc  $\alpha = f(0)$ ,  $\beta = \frac{f(1) - f(-1)}{2}$  et  $\gamma = \frac{f(1) - 2f(0) + f(-1)}{2}$ .

2. On en déduit la méthode de quadrature

$$\int_0^1 f(\tau) d\tau \approx \int_0^1 p(\tau) d\tau = \alpha + \frac{\beta}{2} + \frac{\gamma}{3} = \frac{-f(-1) + 8f(0) + 5f(1)}{12}.$$

3. Soit  $x = m\tau + q$ , alors

$$\int_a^b f(x) dx = m \int_0^1 f(m\tau + q) d\tau \quad \text{avec} \quad \begin{cases} a = q, \\ b = m + q, \end{cases} \quad \text{i.e.} \quad \begin{cases} m = b - a, \\ q = a, \end{cases}$$

d'où le changement de variable  $x = (b - a)\tau + a$ . On en déduit la formule de quadrature

$$\int_a^b f(x) dx = (b - a) \int_0^1 f((b - a)\tau + a) d\tau \approx (b - a) \frac{-f(2a - b) + 8f(a) + 5f(b)}{12}.$$

4. On pose  $a = t_n$  et  $b = t_{n+1}$  d'où la formule de quadrature

$$\int_{t_n}^{t_{n+1}} f(t) dt \approx (t_{n+1} - t_n) \frac{-f(2t_n - t_{n+1}) + 8f(t_n) + 5f(t_{n+1})}{12} = h \frac{-f(t_{n-1}) + 8f(t_n) + 5f(t_{n+1})}{12}.$$

En utilisant la formule de quadrature pour l'intégration de l'EDO  $y'(t) = \varphi(t, y(t))$  entre  $t_n$  et  $t_{n+1}$  on obtient

$$y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} \varphi(t, y(t)) dt \approx h \frac{-\varphi(t_{n-1}, y(t_{n-1})) + 8\varphi(t_n, y(t_n)) + 5\varphi(t_{n+1}, y(t_{n+1}))}{12}.$$

Si on note  $u_n$  une approximation de la solution  $y$  au temps  $t_n$ , on obtient le schéma à deux pas implicite suivant :

$$\begin{cases} u_0 = y(t_0) = y_0, \\ u_1 \text{ à définir,} \\ u_{n+1} = u_n + h \frac{-\varphi(t_{n-1}, u_{n-1}) + 8\varphi(t_n, u_n) + 5\varphi(t_{n+1}, u_{n+1})}{12} & n = 1, 2, \dots, N-1 \end{cases}$$

On peut utiliser une prédiction d'Euler explicite pour initialiser  $u_1$  :

$$\begin{cases} u_0 = y(t_0) = y_0, \\ u_1 = u_0 + h\varphi(t_0, u_0), \\ u_{n+1} = u_n + h \frac{-\varphi(t_{n-1}, u_{n-1}) + 8\varphi(t_n, u_n) + 5\varphi(t_{n+1}, u_{n+1})}{12} & n = 1, 2, \dots, N-1 \end{cases}$$

### Exercice 4.39 (Interpolation, Quadrature et EDO)

1. Soit  $h > 0$  et  $f: [a-h, a+h] \rightarrow \mathbb{R}$  une fonction de classe  $\mathcal{C}^1([a-h, a+h])$ . Écrire le polynôme  $p \in \mathbb{R}_2[x]$  qui interpole  $f$  aux points  $a-h$  et  $a$ , i.e. l'équation de la droite  $p \in \mathbb{R}_2[x]$  qui passe par les deux points  $(a-h, f(a-h))$  et  $(a, f(a))$ .
2. Construire une méthode de quadrature comme suit :

$$\int_a^{a+h} f(x) dx \approx \int_a^{a+h} p(x) dx.$$

NB : on intègre sur  $[a, a+h]$  mais on interpole en  $a-h$  et  $a$ .

3. Considérons le problème de CAUCHY : trouver  $y: [t_0, T] \subset \mathbb{R} \rightarrow \mathbb{R}$  tel que

$$\begin{cases} y'(t) = \varphi(t, y(t)), & \forall t \in [t_0, T], \\ y(t_0) = y_0, \end{cases}$$

dont on suppose l'existence d'une unique solution  $y$ .

On subdivise l'intervalle  $[t_0; T]$  en  $N$  intervalles  $[t_n; t_{n+1}]$  de largeur  $h = \frac{T - t_0}{N}$  avec  $t_n = t_0 + nh$  pour  $n = 0, \dots, N$ . Utiliser la formule obtenue au point 2 pour approcher l'intégrale

$$\int_{t_n}^{t_{n+1}} \varphi(t, y(t)) dt.$$

En déduire un schéma à deux pas explicite pour l'approximation de la solution du problème de CAUCHY.

### Correction

1.  $p(x) = \frac{f(a) - f(a-h)}{a - (a-h)}(x - a) + f(a) = \frac{f(a) - f(a-h)}{h}(x - a) + f(a).$

2. On en déduit la méthode de quadrature

$$\int_a^{a+h} f(x) dx \approx \int_a^{a+h} p(x) dx$$

$$\begin{aligned}
&= \frac{f(a) - f(a-h)}{h} \left[ \frac{(x-a)^2}{2} \right]_a^{a+h} + f(a) [x]_a^{a+h} \\
&= \frac{f(a) - f(a-h)}{2h} ((a+h-a)^2 - (a-a)^2) + f(a)(a+h-a) \\
&= \frac{f(a) - f(a-h)}{2h} h^2 + h f(a) \\
&= h \frac{3f(a) - f(a-h)}{2}.
\end{aligned}$$

3. On pose  $a = t_n$  et  $a + h = t_{n+1}$  d'où la formule de quadrature

$$\int_{t_n}^{t_{n+1}} f(t) dt \approx (t_{n+1} - t_n) \frac{3f(t_n) - f(2t_n - t_{n+1})}{2} = h \frac{3f(t_n) - f(t_{n-1})}{2}.$$

En utilisant la formule de quadrature pour l'intégration de l'EDO  $y'(t) = \varphi(t, y(t))$  entre  $t_n$  et  $t_{n+1}$  on obtient

$$y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} \varphi(t, y(t)) dt \approx h \frac{3\varphi(t_n, y(t_n)) - \varphi(t_{n-1}, y(t_{n-1}))}{2}.$$

Si on note  $u_n$  une approximation de la solution  $y$  au temps  $t_n$ , on obtient le schéma à deux pas implicite suivant :

$$\begin{cases} u_0 = y(t_0) = y_0, \\ u_1 \text{ à définir,} \\ u_{n+1} = u_n + h \frac{3\varphi(t_{n-1}, u_{n-1}) - \varphi(t_n, u_n)}{2} \quad n = 1, 2, \dots, N-1 \end{cases}$$

On peut utiliser une prédition d'Euler explicite pour initialiser  $u_1$  :

$$\begin{cases} u_0 = y(t_0) = y_0, \\ u_1 = u_0 + h\varphi(t_0, u_0), \\ u_{n+1} = u_n + h \frac{3\varphi(t_{n-1}, u_{n-1}) - \varphi(t_n, u_n)}{2} \quad n = 1, 2, \dots, N-1 \end{cases}$$

# Chapitre 5

## Systèmes linéaires

### 5.1 Rappels d'algèbre linéaire

#### Définition 5.1 (Système linéaire)

Soit  $n, p \geq 1$  des entiers. Un SYSTÈME LINÉAIRE  $n \times p$  est un ensemble de  $n$  équations linéaires à  $p$  inconnues de la forme

$$(S) \quad \begin{cases} a_{11}x_1 + \dots + a_{1p}x_p = b_1, \\ \vdots \\ a_{n1}x_1 + \dots + a_{np}x_p = b_n. \end{cases}$$

- ★ Les COEFFICIENTS  $a_{ij}$  et les SECONDES MEMBRES  $b_i$  sont des éléments donnés de  $\mathbb{K}$ .
- ★ Les INCONNUES  $x_1, x_2, \dots, x_p$  sont à chercher dans  $\mathbb{K}$ .
- ★ Une SOLUTION de  $(S)$  est un  $p$ -uplet  $(x_1, x_2, \dots, x_p)$  qui vérifie simultanément les  $n$  équations de  $(S)$ . Résoudre  $(S)$  signifie chercher toutes les solutions.
- ★ Un système est IMPOSSIBLE, ou incompatible, s'il n'admet pas de solution.  
Un système est POSSIBLE, ou compatible, s'il admet une ou plusieurs solutions.
- ★ Le SYSTÈME HOMOGÈNE associé à  $(S)$  est le système obtenu en remplaçant les  $b_i$  par 0.
- ★ Deux systèmes sont ÉQUIVALENTS s'ils admettent les mêmes solutions.

#### Écriture matricielle Si on note

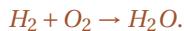
$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \quad \mathbb{A} = \begin{pmatrix} a_{11} & \dots & a_{1p} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{np} \end{pmatrix}$$

le système  $(S)$  est équivalent à l'écriture matricielle  $\mathbb{A}\mathbf{x} = \mathbf{b}$ .

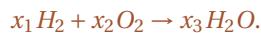
Il est fréquent, dans toutes les disciplines scientifiques, de devoir résoudre des systèmes linéaires de la forme  $\mathbb{A}\mathbf{x} = \mathbf{b}$ , où  $\mathbb{A}$  est une matrice carrée de dimension  $n \times n$  dont les éléments  $a_{ij}$  sont réels ou complexes et  $\mathbf{x}$  et  $\mathbf{b}$  sont des vecteurs colonnes de dimension  $n$ , où  $\mathbf{x}$  est l'inconnue et  $\mathbf{b}$  un vecteur donné.

 **Équilibrage de réactions chimiques** Du point de vue mathématique, équilibrer une réaction chimique signifie trouver des coefficients (dans  $\mathbb{N}$  ou  $\mathbb{Q}$ ), appelés coefficients stœchiométriques, qui satisfont certaines contraintes comme la conservation du nombre d'atomes, ou la conservation du nombre d'électrons (pour les réactions red-ox), ou la conservation de la charge (pour les réactions écrites sous forme ionique). Toutes ces contraintes dépendent linéairement des coefficients stœchiométriques, ce qui amène tout naturellement à l'écriture d'un système linéaire.

Par exemple, considérons la réaction



Notons  $x_1$ ,  $x_2$  et  $x_3$  les coefficients stœchiométriques



Les contraintes sont :

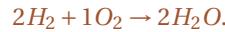
1. la conservation du nombre d'atomes d'hydrogène :  $2x_1 = 2x_3$ ,
2. la conservation du nombre d'atomes d'oxygène :  $x_2 = x_3$ .

On note qu'on a 3 inconnues mais seulement 2 équations linéairement indépendantes ; en effet, les coefficients stœchiométriques ne définissent pas des quantités absolues mais seulement les rapports entre les différents éléments. Par conséquent, si  $(x_1, x_2, x_3)$  équilibre la réaction, alors tous les multiples entiers de  $(x_1, x_2, x_3)$  équilibreront aussi la réaction.

Pour résoudre le problème sans paramètres, fixons arbitrairement un des coefficients, par exemple  $x_3 = 1$ . On doit alors résoudre le système linéaire

$$\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

On trouve alors  $x_1 = 1$  et  $x_2 = 1/2$ . Si nous voulons des coefficients stœchiométriques entiers, il suffit de multiplier tous les coefficients par 2 et on a ainsi



### Partage de secrets *Comment envoyer un message secret avec plusieurs espions sans pour autant que ceux-ci ne connaissent le contenu du message envoyé ?*

Typiquement, un message à envoyer est un nombre entier (car, par codage, on peut remplacer un texte quelconque par un nombre). Imaginons donc que l'on désire envoyer le nombre  $n$ . Considérons un polynôme de degré  $k$ , par exemple à coefficients entiers,  $P(X) = a_k X^k + \dots + a_1 X + n$  dont le terme indépendant vaut exactement  $n$ . En particulier, on a  $P(0) = n$ . Un corollaire du théorème fondamental de l'algèbre stipule que si deux polynômes de degré  $k$  sont égaux en  $k+1$  points, alors ils sont égaux. Autrement dit, le polynôme  $P$  est complètement caractérisé par les valeurs qu'il prend par exemple aux entiers  $1, 2, \dots, k+1$ . On engage alors  $k+1$  espions (voire un peu plus, si certains étaient capturés par les «ennemis»). On donne au  $i$ -ème espion le nombre  $P(i)$ . Les espions se dispersent (par exemple, pour passer les lignes ennemis). Une fois que  $k+1$  espions sont arrivés à destination, il est aisément de reconstituer le polynôme (on a *un système de  $k+1$  équations linéaires* pour retrouver les  $k+1$  coefficients de  $P$ ) et ainsi retrouver la valeur secrète  $n$ . Si un espion est capturé et qu'il parle, les ennemis auront à leur disposition un des  $P(i)$ , mais cela ne leur permet nullement de retrouver  $n$ . De même, si un espion étaient en fait un agent double, connaître  $P(i)$  seul ne sert à rien.

Source : <http://michelrigo.wordpress.com/2010/01/30/partage-de-secrets-et-tfa/>

## 5.2 Méthodes de résolution analytiques

### Définition 5.2 (Systèmes triangulaires)

On dit qu'une matrice carrée  $\mathbb{A} = (a_{ij})_{1 \leq i, j \leq n}$  est TRIANGULAIRE SUPÉRIEURE (respectivement triangulaire INFÉRIEURE) si  $i > j \Rightarrow a_{ij} = 0$  (resp. si  $i < j \Rightarrow a_{ij} = 0$ ).

Si la matrice est triangulaire supérieure (resp. triangulaire inférieure), on dira que le système linéaire est un système triangulaire supérieur (resp. triangulaire inférieur).

Pour résoudre le système triangulaire  $\mathbb{A}\mathbf{x} = \mathbf{b}$ ,

- \* si  $\mathbb{A}$  est une matrice triangulaire inférieure, on a  $x_1 = \frac{b_1}{a_{11}}$  et on déduit les inconnues  $x_2, x_3, \dots, x_n$  grâce à la relation  $x_i = \frac{1}{a_{ii}} \left( b_i - \sum_{j=1}^{i-1} a_{ij} x_j \right)$ ;
- \* si  $\mathbb{A}$  est une matrice triangulaire supérieure on a  $x_n = \frac{b_n}{a_{nn}}$  et on déduit les inconnues  $x_{n-1}, x_{n-2}, \dots, x_1$  grâce à la relation  $x_i = \frac{1}{a_{ii}} \left( b_i - \sum_{j=i+1}^n a_{ij} x_j \right)$ .

### Définition 5.3 (Système échelonné)

Un système (S) est EN ESCALIER, ou ÉCHELONNÉ, si le nombre de premiers coefficients nuls successifs de chaque équation est strictement croissant.

Autrement dit, un système est échelonné si les coefficients non nuls des équations se présentent avec une sorte d'escalier à marches de longueurs variables marquant la séparation entre une zone composée uniquement de zéros et une zone où les lignes situées à droite de l'escalier commencent par des termes non nuls, comme dans l'exemple suivant de 5 équations à 6 inconnues :

$$\left\{ \begin{array}{rcl} 5x_1 - x_2 - x_3 + 2x_4 & & +x_6 = b_1 \\ & 3x_3 - x_4 & +2x_5 = b_2 \\ & & -x_5 + x_6 = b_3 \\ & & 5x_6 = b_4 \\ & & 0 = b_5 \end{array} \right.$$

 **Réduction** Quand un système contient une équation du type

$$0x_1 + \cdots + 0x_p = b,$$

- \* si  $b \neq 0$  le système est impossible,
- \* si  $b = 0$ , on peut supprimer cette équation, ce qui conduit à un système équivalent à (S) dit **SYSTÈME RÉDUIT**.

 **Définition 5.4 (Matrice augmentée)**

Si on ajoute le vecteur-colonne des seconds membres **b** à la matrice des coefficients **A**, on obtient ce qu'on appelle la matrice augmentée que l'on note  $[A|b]$ .

 **Méthode de GAUSS** Soit  $\mathbb{A} = (a_{ij})_{\substack{1 \leq i \leq n \\ 1 \leq j \leq p}}$  la matrice des coefficients du système (S) et  $[A|b]$  la matrice augmentée.

La méthode de GAUSS comporte  $n - 1$  étapes : à chaque étape  $j$  on fait apparaître des 0 sur la colonne  $j$  pour les lignes  $i > j$  par des opérations élémentaires sur les lignes.

*Étape j* : en permutant éventuellement deux lignes de la matrice augmentée (*i.e.* deux équations du système linéaire), on peut supposer  $a_{jj} \neq 0$  (appelé pivot de l'étape  $j$ ). On transforme alors toutes les lignes  $L_i$  avec  $i > j$  selon la règle :

$$L_i \leftarrow L_i - \frac{a_{ij}}{a_{jj}} L_j,$$

ainsi on fait apparaître des 0 sur la colonne  $j$  pour les lignes  $i > j$  (*i.e.* on élimine l'inconnue  $x_j$  dans chaque lignes  $L_i$  du système linéaire).

En réitérant le procédé pour  $i$  de 1 à  $n - 1$ , on aboutit à un système échelonné.

 **EXEMPLE**

Soit le système linéaire

$$\begin{cases} x_1 + 2x_2 + 3x_3 + 4x_4 = 1, \\ 2x_1 + 3x_2 + 4x_3 + x_4 = 2, \\ 3x_1 + 4x_2 + x_3 + 2x_4 = 3, \\ 4x_1 + x_2 + 2x_3 + 3x_4 = 4. \end{cases}$$

1. Résolution par la méthode du pivot de GAUSS :

$$\begin{array}{l} \left\{ \begin{array}{l} x_1 + 2x_2 + 3x_3 + 4x_4 = 1 \\ 2x_1 + 3x_2 + 4x_3 + x_4 = 2 \\ 3x_1 + 4x_2 + x_3 + 2x_4 = 3 \\ 4x_1 + x_2 + 2x_3 + 3x_4 = 4 \end{array} \right. \xrightarrow{\substack{L_2 \leftarrow L_2 - 2L_1 \\ L_3 \leftarrow L_3 - 3L_1 \\ L_4 \leftarrow L_4 - 4L_1}} \left\{ \begin{array}{l} x_1 + 2x_2 + 3x_3 + 4x_4 = 1 \\ -x_2 - 2x_3 - 7x_4 = 0 \\ -2x_2 - 8x_3 - 10x_4 = 0 \\ -7x_2 - 10x_3 - 13x_4 = 0 \end{array} \right. \\ \text{Étape 1} \\ \xrightarrow{\substack{L_3 \leftarrow L_3 - 2L_2 \\ L_4 \leftarrow L_4 - 7L_2}} \left\{ \begin{array}{l} x_1 + 2x_2 + 3x_3 + 4x_4 = 1 \\ -x_2 - 2x_3 - 7x_4 = 0 \\ -4x_3 + 4x_4 = 0 \\ 4x_3 + 36x_4 = 0 \end{array} \right. \\ \text{Étape 2} \\ \xrightarrow{\substack{L_4 \leftarrow L_4 + L_3}} \left\{ \begin{array}{l} x_1 + 2x_2 + 3x_3 + 4x_4 = 1 \\ -x_2 - 2x_3 - 7x_4 = 0 \\ -4x_3 + 4x_4 = 0 \\ 40x_4 = 0 \end{array} \right. \\ \text{Étape 3} \end{array}$$

donc, en résolvant le système triangulaire supérieur obtenu, on obtient

$$x_4 = 0, \quad x_3 = 0, \quad x_2 = 0, \quad x_1 = 1.$$

2. Résolution par la méthode du pivot de GAUSS en écriture matricielle :

$$\begin{array}{l} [\mathbb{A}|b] = \left( \begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 2 & 3 & 4 & 1 & 2 \\ 3 & 4 & 1 & 2 & 3 \\ 4 & 1 & 2 & 3 & 4 \end{array} \right) \xrightarrow{\substack{L_2 \leftarrow L_2 - 2L_1 \\ L_3 \leftarrow L_3 - 3L_1 \\ L_4 \leftarrow L_4 - 4L_1}} \left( \begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 0 & -1 & -2 & -7 & 0 \\ 0 & -2 & -8 & -10 & 0 \\ 0 & -7 & -10 & -13 & 0 \end{array} \right) \\ \text{Étape 1} \\ \xrightarrow{\substack{L_3 \leftarrow L_3 - 2L_2 \\ L_4 \leftarrow L_4 - 7L_2}} \left( \begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 0 & -1 & -2 & -7 & 0 \\ 0 & 0 & -4 & 4 & 0 \\ 0 & 0 & 4 & 36 & 0 \end{array} \right) \xrightarrow{\substack{L_4 \leftarrow L_4 + L_3}} \left( \begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 0 & -1 & -2 & -7 & 0 \\ 0 & 0 & -4 & 4 & 0 \\ 0 & 0 & 0 & 40 & 0 \end{array} \right) \\ \text{Étape 2} \\ \text{Étape 3} \end{array}$$

donc

$$x_4 = 0, \quad x_3 = 0, \quad x_2 = 0, \quad x_1 = 1.$$

**Méthode de GAUSS-JORDAN** Soit  $\mathbb{A} = (a_{ij})_{\substack{1 \leq i \leq n \\ 1 \leq j \leq p}}$  la matrice des coefficients du système (S) et  $[\mathbb{A}|\mathbf{b}]$  la matrice augmentée.

Dans cette variante de la méthode du pivot de GAUSS, à chaque étape on fait apparaître des zéros à la fois au-dessus et en-dessous du pivot.

Étape  $j$  : en permutant éventuellement deux lignes de la matrice augmentée, on peut supposer  $a_{jj} \neq 0$ . On transforme alors toutes les lignes  $L_i$  avec  $i \neq j$  selon la règle

$$L_i \leftarrow L_i - \frac{a_{ij}}{a_{jj}} L_j$$

ainsi on élimine l'inconnue  $x_j$  dans toutes les lignes  $L_i$ .

En réitérant le procédé pour  $i$  de 1 à  $n$ , on aboutit à un système diagonal.

### EXEMPLE

Résoudre le système linéaire

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \\ 3 & 4 & 1 & 2 \\ 4 & 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$$

par la méthode de GAUSS-JORDAN.

$$[\mathbb{A}|\mathbf{b}] = \left( \begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 2 & 3 & 4 & 1 & 2 \\ 3 & 4 & 1 & 2 & 3 \\ 4 & 1 & 2 & 3 & 4 \end{array} \right) \xrightarrow{\substack{L_2 \leftarrow L_2 - 2L_1 \\ L_3 \leftarrow L_3 - 3L_1 \\ L_4 \leftarrow L_4 - 4L_1 \\ \text{Étape 1}}} \left( \begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 0 & -1 & -2 & -7 & 0 \\ 0 & -2 & -8 & -10 & 0 \\ 0 & -7 & -10 & -13 & 0 \end{array} \right) \xrightarrow{\substack{L_1 \leftarrow L_1 + 2L_2 \\ L_3 \leftarrow L_3 - 2L_2 \\ L_4 \leftarrow L_4 - 7L_2 \\ \text{Étape 2}}} \left( \begin{array}{cccc|c} 1 & 0 & -1 & -10 & 1 \\ 0 & -1 & -2 & -7 & 0 \\ 0 & 0 & -4 & 4 & 0 \\ 0 & 0 & 4 & 36 & 0 \end{array} \right) \\ \xrightarrow{\substack{L_1 \leftarrow L_1 - L_3/4 \\ L_2 \leftarrow L_2 - L_3/2 \\ L_4 \leftarrow L_4 + L_3 \\ \text{Étape 3}}} \left( \begin{array}{cccc|c} 1 & 0 & 0 & 4 & 1 \\ 0 & -1 & 0 & -7 & 0 \\ 0 & 0 & -4 & 4 & 0 \\ 0 & 0 & 0 & 40 & 0 \end{array} \right) \xrightarrow{\substack{L_1 \leftarrow L_1 + 11L_4/40 \\ L_2 \leftarrow L_2 + 9L_4/40 \\ L_3 \leftarrow L_3 + 4L_4/40 \\ \text{Étape 4}}} \left( \begin{array}{cccc|c} 1 & 0 & 0 & 0 & 1 \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & -4 & 0 & 0 \\ 0 & 0 & 0 & 40 & 0 \end{array} \right)$$

donc

$$x_1 = 1, \quad x_2 = 0, \quad x_3 = 0, \quad x_4 = 0.$$

### Définition 5.5 (Rang)

Le nombre d'équations non triviales du système réduit en escalier obtenu par la méthode de GAUSS est le RANG  $r$  DE LA MATRICE  $\mathbb{A}$ , OU DU SYSTÈME (S).

### Théorème 5.6

Un système carré  $\mathbb{A}\mathbf{x} = \mathbf{b}$  de  $n$  équations à  $n$  inconnues est compatible si et seulement si  $\text{rg}(\mathbb{A}) = \text{rg}([\mathbb{A}|\mathbf{b}])$ .

1. Si  $\text{rg}(\mathbb{A}) = n$  (i.e. si  $\det(\mathbb{A}) \neq 0$ ) alors  $\text{rg}(\mathbb{A}) = \text{rg}([\mathbb{A}|\mathbf{b}])$  et la solution est unique.
2. Si  $\text{rg}(\mathbb{A}) = \text{rg}([\mathbb{A}|\mathbf{b}]) < n$  il y a une infinité de solutions.
3. Si  $\text{rg}(\mathbb{A}) \neq \text{rg}([\mathbb{A}|\mathbf{b}])$  il n'y a pas de solution.

### EXEMPLE

On veut résoudre les systèmes linéaires suivants de 2 équations et 2 inconnues :

$$\textcircled{1} \quad \begin{cases} x + y = 1 \\ x - y = 1 \end{cases}$$

$$\textcircled{2} \quad \begin{cases} x + y = 1 \\ 2x + 2y = 2 \end{cases}$$

$$\textcircled{3} \quad \begin{cases} x + y = 1 \\ 2x + 2y = 1 \end{cases}$$

Les matrices augmentées associées à chaque système sont

$$\textcircled{1} \quad [\mathbb{A}|\mathbf{b}] = \left[ \begin{array}{cc|c} 1 & 1 & 1 \\ 1 & -1 & 1 \end{array} \right]$$

$$\textcircled{2} \quad [\mathbb{A}|\mathbf{b}] = \left[ \begin{array}{cc|c} 1 & 1 & 1 \\ 2 & 2 & 2 \end{array} \right]$$

$$\textcircled{3} \quad [\mathbb{A}|\mathbf{b}] = \left[ \begin{array}{cc|c} 1 & 1 & 1 \\ 2 & 2 & 1 \end{array} \right]$$

et on a

①  $\text{rg}(\mathbb{A}) = \text{rg}([\mathbb{A}|\mathbf{b}]) = 2$  donc il existe une et une seule solution. En effet,

$$\begin{cases} x + y = 1 \\ x - y = 1 \end{cases} \xrightarrow{L_2 \leftarrow L_2 - L_1} \begin{cases} x + y = 1 \\ -2y = 0 \end{cases}$$

ainsi la solution est  $y = 0$  et  $x = 1$  ;

②  $\text{rg}(\mathbb{A}) = \text{rg}([\mathbb{A}|\mathbf{b}]) = 1$  donc il existe une infinité de solutions. En effet,

$$\begin{cases} x + y = 1 \\ 2x + 2y = 2 \end{cases} \xrightarrow{L_2 \leftarrow L_2 - 2L_1} \begin{cases} x + y = 1 \\ 0 = 0 \end{cases}$$

ainsi la solution est  $y = \kappa$  et  $x = 1 - \kappa$  pour tout  $\kappa \in \mathbb{R}$  ;

③  $\text{rg}(\mathbb{A}) = 1$  et  $\text{rg}([\mathbb{A}|\mathbf{b}]) = 2$  donc il n'y a pas de solution. En effet

$$\begin{cases} x + y = 1 \\ 2x + 2y = 1 \end{cases} \xrightarrow{L_2 \leftarrow L_2 - 2L_1} \begin{cases} x + y = 1 \\ 0 = -1 \end{cases}$$

et la dernière équation est impossible.

### Définition 5.7 (Système de CRAMER)

Un SYSTÈME est dit DE CRAMER s'il a une solution, et une seule.

### Propriété 5.8

Considérons un système carré d'ordre  $n$  à coefficients réels. Le système est de CRAMER si une des conditions équivalentes suivantes est remplie :

1.  $\mathbb{A}$  est inversible ;
2.  $\text{rg}(\mathbb{A}) = n$  ;
3. le système homogène  $\mathbb{A}\mathbf{x} = \mathbf{0}$  admet seulement la solution nulle.

 **Méthode de CRAMER** La solution d'un système de CRAMER d'écriture matricielle  $\mathbb{A}\mathbf{x} = \mathbf{b}$  est donnée par

$$x_j = \frac{\det(\mathbb{A}_j)}{\det(\mathbb{A})}, \quad 1 \leq j \leq n$$

où  $\mathbb{A}_j$  est la matrice obtenue à partir de  $\mathbb{A}$  en remplaçant la  $j$ -ème colonne par la colonne des seconds membres  $\mathbf{b}$ .

Cette formule est cependant d'une utilité pratique limitée à cause du calcul des déterminants qui est très coûteux.

### EXEMPLE (SYSTÈME D'ORDRE 2)

On veut résoudre le système linéaire

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

par la méthode de CRAMER. On a

$$\begin{aligned} \mathbb{A} &= \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, & \det(\mathbb{A}) &= a_{11}a_{22} - a_{12}a_{21}, \\ \mathbb{A}_1 &= \begin{pmatrix} b_1 & a_{12} \\ b_2 & a_{22} \end{pmatrix}, & \det(\mathbb{A}_1) &= b_1a_{22} - a_{12}b_2, \\ \mathbb{A}_2 &= \begin{pmatrix} a_{11} & b_1 \\ a_{21} & b_2 \end{pmatrix}, & \det(\mathbb{A}_2) &= a_{11}b_2 - b_1a_{21}, \end{aligned}$$

donc

$$x_1 = \frac{b_1a_{22} - a_{12}b_2}{a_{11}a_{22} - a_{12}a_{21}}, \quad x_2 = \frac{a_{11}b_2 - b_1a_{21}}{a_{11}a_{22} - a_{12}a_{21}}.$$

 EXEMPLE

On veut résoudre le système linéaire

$$\begin{pmatrix} 1 & -1 & 2 \\ 2 & 1 & 0 \\ 3 & 2 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix}$$

par la méthode de CRAMER. On a

$$\mathbb{A} = \begin{pmatrix} 1 & -1 & 2 \\ 2 & 1 & 0 \\ 3 & 2 & 0 \end{pmatrix}, \quad \det(\mathbb{A}) = 2,$$

$$\mathbb{A}_1 = \begin{pmatrix} 2 & -1 & 2 \\ -1 & 1 & 0 \\ 1 & 2 & 0 \end{pmatrix}, \quad \det(\mathbb{A}_1) = -6,$$

$$\mathbb{A}_2 = \begin{pmatrix} 1 & 2 & 2 \\ 2 & -1 & 0 \\ 3 & 1 & 0 \end{pmatrix}, \quad \det(\mathbb{A}_2) = 10,$$

$$\mathbb{A}_3 = \begin{pmatrix} 1 & -1 & 2 \\ 2 & 1 & -1 \\ 3 & 2 & 1 \end{pmatrix}, \quad \det(\mathbb{A}_3) = 10,$$

donc

$$x = \frac{-6}{2} = -3, \quad y = \frac{10}{2} = 5, \quad z = \frac{10}{2} = 5.$$



**Définition 5.9 (Cofacteur & comatrice)**

Soit  $\mathbb{A}$  une matrice carrée d'ordre  $n$ . Étant donné un couple  $(i, j)$  d'entiers,  $1 \leq i, j \leq n$ , on note  $\mathbb{A}_{ij}$  la matrice carrée d'ordre  $n-1$  obtenue en supprimant la  $i$ -ème ligne et la  $j$ -ème colonne de  $\mathbb{A}$ . On appelle COFACTEUR de l'élément  $a_{ij}$  le nombre  $(-1)^{i+j} \det(\mathbb{A}_{ij})$ . On appelle COMATRICE de  $\mathbb{A}$  la matrice constituée des cofacteurs de  $\mathbb{A}$ .



**EXEMPLE**

Soit  $\mathbb{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ . Alors la matrice des cofacteurs de  $\mathbb{A}$  est la matrice  $\begin{pmatrix} d & -c \\ -b & a \end{pmatrix}$ .



**Calcul de  $\mathbb{A}^{-1}$**

$\mathbb{A}$  étant inversible, pour obtenir  $\mathbb{A}^{-1}$  il suffit de résoudre le système  $\mathbb{A}\mathbf{x} = \mathbf{b}$  qui admet pour solution  $\mathbf{x} = \mathbb{A}^{-1}\mathbf{b}$ . On peut alors calculer  $\mathbb{A}^{-1}$  en résolvant  $n$  systèmes linéaires de termes sources  $(1, 0, 0, \dots, 0), (0, 1, 0, \dots, 0), \dots, (0, 0, 0, \dots, 1)$ . Les méthodes suivantes résolvent ces  $n$  systèmes linéaires simultanément.

Première méthode.

1. On calcul la matrice des cofacteurs des éléments de  $\mathbb{A}$ , appelée comatrice de  $\mathbb{A}$  ;
2. on transpose la comatrice de  $\mathbb{A}$  ;
3. on divise par  $\det(\mathbb{A})$ .

Cette méthode est quasi-impraticable dès que  $n > 3$ .

Deuxième méthode.

La matrice  $\mathbb{A}$  est inversible si et seulement si on obtient par opérations élémentaires sur les lignes de  $\mathbb{A}$  une matrice triangulaire sans zéros sur la diagonale ; non inversible si et seulement si on obtient une matrice triangulaire avec un zéro sur la diagonale. Si  $\mathbb{A}$  est inversible, on effectue les mêmes opérations sur la matrice  $[\mathbb{A} | \mathbb{I}_n]$  jusqu'à obtenir  $[\mathbb{I}_n | \mathbb{A}^{-1}]$ :

$$[\mathbb{A} | \mathbb{I}_n] \xrightarrow{\text{Opérations élémentaires}} [\mathbb{I}_n | \mathbb{A}^{-1}].$$



**EXEMPLE**

Soit  $\mathbb{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$  avec  $\det(\mathbb{A}) = ad - bc \neq 0$ .

Première méthode : on a déjà calculé le déterminant de cette matrice ainsi que la matrice des cofacteurs, il suffit alors de calculer la transposée et on obtient

$$\mathbb{A}^{-1} = \frac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

Deuxième méthode : on parvient au même résultat par transformations élémentaires :

$$\begin{aligned} [\mathbb{A}|\mathbb{I}_2] &= \left( \begin{array}{cc|cc} a & b & 1 & 0 \\ c & d & 0 & 1 \end{array} \right) \xrightarrow{L_2 \leftarrow L_2 - \frac{c}{a}L_1} \left( \begin{array}{cc|cc} a & b & 1 & 0 \\ 0 & d - \frac{c}{a}b & -\frac{c}{a} & 1 \end{array} \right) \\ &\xrightarrow{L_1 \leftarrow L_1 - \frac{b}{d-\frac{c}{a}b}L_2} \left( \begin{array}{cc|cc} a & 0 & 1 + \frac{bc}{ad-bc} & -\frac{ab}{ad-bc} \\ 0 & d - \frac{c}{a}b & -\frac{c}{a} & 1 \end{array} \right) \\ &\xrightarrow{L_2 \leftarrow L_2 - \frac{1}{d-\frac{c}{a}b}L_1} \left( \begin{array}{cc|cc} 1 & 0 & \frac{1}{a} + \frac{bc}{a(ad-bc)} & -\frac{ab}{ad-bc} \\ 0 & 1 & -\frac{c}{ad-cb} & \frac{a}{ad-cb} \end{array} \right) = \left( \begin{array}{cc|cc} 1 & 0 & \frac{d}{ad-bc} & -\frac{b}{ad-bc} \\ 0 & 1 & -\frac{c}{ad-cb} & \frac{a}{ad-cb} \end{array} \right) \end{aligned}$$

### EXEMPLE

Calculer l'inverse de la matrice

$$\mathbb{A} = \begin{pmatrix} 1 & 1 & -1 \\ -1 & 1 & 1 \\ 1 & -1 & 1 \end{pmatrix}.$$

### Première méthode.

- On calcule la matrice des cofacteurs des éléments de  $\mathbb{A}$ , appelée comatrice de  $\mathbb{A}$  :

$$\text{comatrice} = \begin{pmatrix} (-1)^{1+1} \begin{vmatrix} 1 & 1 \\ -1 & 1 \end{vmatrix} & (-1)^{1+2} \begin{vmatrix} -1 & 1 \\ 1 & 1 \end{vmatrix} & (-1)^{1+3} \begin{vmatrix} -1 & 1 \\ 1 & -1 \end{vmatrix} \\ (-1)^{2+1} \begin{vmatrix} 1 & -1 \\ -1 & 1 \end{vmatrix} & (-1)^{2+2} \begin{vmatrix} 1 & -1 \\ 1 & 1 \end{vmatrix} & (-1)^{2+3} \begin{vmatrix} 1 & 1 \\ 1 & -1 \end{vmatrix} \\ (-1)^{3+1} \begin{vmatrix} 1 & -1 \\ 1 & 1 \end{vmatrix} & (-1)^{3+2} \begin{vmatrix} 1 & -1 \\ -1 & 1 \end{vmatrix} & (-1)^{3+3} \begin{vmatrix} 1 & 1 \\ -1 & 1 \end{vmatrix} \end{pmatrix} = \begin{pmatrix} 2 & 2 & 0 \\ 0 & 2 & 2 \\ 2 & 0 & 2 \end{pmatrix};$$

- on transpose la comatrice de  $\mathbb{A}$  :

$$\text{comatrice}^T = \begin{pmatrix} 2 & 0 & 2 \\ 2 & 2 & 0 \\ 0 & 2 & 2 \end{pmatrix};$$

- on divise par  $\det(\mathbb{A})$  :

$$\mathbb{A}^{-1} = \frac{1}{4} \begin{pmatrix} 2 & 0 & 2 \\ 2 & 2 & 0 \\ 0 & 2 & 2 \end{pmatrix} = \begin{pmatrix} 1/2 & 0 & 1/2 \\ 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \end{pmatrix}.$$

### Deuxième méthode.

$$\begin{aligned} [\mathbb{A}|\mathbb{I}_3] &= \left( \begin{array}{ccc|ccc} 1 & 1 & -1 & 1 & 0 & 0 \\ -1 & 1 & 1 & 0 & 1 & 0 \\ 1 & -1 & 1 & 0 & 0 & 1 \end{array} \right) \xrightarrow{L_2 \leftarrow L_2 + L_1} \left( \begin{array}{ccc|ccc} 1 & 1 & -1 & 1 & 0 & 0 \\ 0 & 2 & 0 & 1 & 1 & 0 \\ 1 & -1 & 1 & 0 & 0 & 1 \end{array} \right) \\ &\xrightarrow{L_3 \leftarrow L_3 - L_1} \left( \begin{array}{ccc|ccc} 1 & 1 & -1 & 1 & 0 & 0 \\ 0 & 2 & 0 & 1 & 1 & 0 \\ 0 & -2 & 2 & -1 & 0 & 1 \end{array} \right) \xrightarrow{L_2 \leftarrow L_2/2} \left( \begin{array}{ccc|ccc} 1 & 1 & -1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1/2 & 1/2 & 0 \\ 0 & -2 & 2 & -1 & 0 & 1 \end{array} \right) \\ &\xrightarrow{L_3 \leftarrow L_3 + 2L_2} \left( \begin{array}{ccc|ccc} 1 & 1 & -1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 2 & 0 & 1 & 1 \end{array} \right) \xrightarrow{L_1 \leftarrow L_1 - L_2} \left( \begin{array}{ccc|ccc} 1 & 0 & -1 & 1/2 & -1/2 & 0 \\ 0 & 1 & 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 2 & 0 & 1 & 1 \end{array} \right) \\ &\xrightarrow{L_3 \leftarrow L_3/2} \left( \begin{array}{ccc|ccc} 1 & 0 & -1 & 1/2 & -1/2 & 0 \\ 0 & 1 & 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 1 & 0 & 1/2 & 1/2 \end{array} \right) \xrightarrow{L_1 \leftarrow L_1 + L_3} \left( \begin{array}{ccc|ccc} 1 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 1 & 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 1 & 0 & 1/2 & 1/2 \end{array} \right) = [\mathbb{I}_3 | \mathbb{A}^{-1}]. \end{aligned}$$

### 5.2.1 Systèmes rectangulaires

#### Théorème 5.10

Un système  $\mathbb{A}\mathbf{x} = \mathbf{b}$  de  $m$  équations à  $n$  inconnues est compatible si et seulement si  $\text{rg}(\mathbb{A}) = \text{rg}([\mathbb{A}|\mathbf{b}])$ .

1. Si le système a  $n$  équations et  $n$  inconnues, la matrice  $\mathbb{A}$  est carrée d'ordre  $n$  et 3 situations peuvent se présenter :
  - 1.1. Si  $\text{rg}(\mathbb{A}) = n$  (i.e. si  $\det(\mathbb{A}) \neq 0$ ) alors  $\text{rg}(\mathbb{A}) = \text{rg}([\mathbb{A}|\mathbf{b}])$  et la solution est unique.
  - 1.2. Si  $\text{rg}(\mathbb{A}) = \text{rg}([\mathbb{A}|\mathbf{b}]) < n$  il y a une infinité de solutions.
  - 1.3. Si  $\text{rg}(\mathbb{A}) \neq \text{rg}([\mathbb{A}|\mathbf{b}])$  il n'y a pas de solution.
2. Si le système a  $m$  équations et  $n$  inconnues avec  $m > n$  alors 3 situations peuvent se présenter :
  - 2.1. Si  $\text{rg}(\mathbb{A}) = \text{rg}([\mathbb{A}|\mathbf{b}]) = n$  la solution est unique.
  - 2.2. Si  $\text{rg}(\mathbb{A}) = \text{rg}([\mathbb{A}|\mathbf{b}]) < n$  il y a une infinité de solutions.
  - 2.3. Si  $\text{rg}(\mathbb{A}) \neq \text{rg}([\mathbb{A}|\mathbf{b}])$  il n'y a pas de solution.
3. Si le système a  $m$  équations et  $n$  inconnues avec  $m < n$  alors 2 situations peuvent se présenter :
  - 3.1. Si  $\text{rg}(\mathbb{A}) = \text{rg}([\mathbb{A}|\mathbf{b}]) \leq m < n$  il y a une infinité de solutions.
  - 3.2. Si  $\text{rg}(\mathbb{A}) \neq \text{rg}([\mathbb{A}|\mathbf{b}])$  il n'y a pas de solution.

#### Remarque

Soit  $\mathbb{A} \in \mathcal{M}_{n,p}$  la matrice des coefficients du système (S). Alors

$$\begin{aligned} 0 \leq \text{rg}(\mathbb{A}) &\leq \min \{ n, p \} \\ \text{rg}(\mathbb{A}) &\leq \text{rg}([\mathbb{A}|\mathbf{b}]) \leq \min \{ n, p + 1 \}. \end{aligned}$$

#### EXEMPLE

1.  $n$  équations et  $n$  inconnues :

- 1.1.  $\mathbb{A} = \begin{pmatrix} 1 & 2 & 3 \\ 1 & -3 & -7 \\ -6 & 4 & -2 \end{pmatrix}, \mathbf{b} = \begin{pmatrix} 12 \\ -26 \\ -4 \end{pmatrix}$ . On a  $\text{rg}(\mathbb{A}) = 3$  (car  $\det(\mathbb{A}) \neq 0$ ) donc  $\text{rg}(\mathbb{A}) = \text{rg}([\mathbb{A}|\mathbf{b}])$  et la solution est unique.
- 1.2.  $\mathbb{A} = \begin{pmatrix} 1 & 2 & 3 \\ 1 & -3 & -7 \\ 3 & -2 & -7 \end{pmatrix}, \mathbf{b} = \begin{pmatrix} 14 \\ -26 \\ -22 \end{pmatrix}$ . On a  $\text{rg}(\mathbb{A}) = \text{rg}([\mathbb{A}|\mathbf{b}]) = 2 < 3$  donc il y a une infinité de solutions.
- 1.3.  $\mathbb{A} = \begin{pmatrix} 1 & 2 & 3 \\ 1 & -3 & -7 \\ 3 & -2 & -7 \end{pmatrix}, \mathbf{b} = \begin{pmatrix} 14 \\ -26 \\ -20 \end{pmatrix}$ . On a  $\text{rg}(\mathbb{A}) = 2 \neq \text{rg}([\mathbb{A}|\mathbf{b}]) = 3$  donc il n'y a pas de solution.

2.  $m$  équations et  $n$  inconnues avec  $m > n$  :

- 2.1.  $\mathbb{A} = \begin{pmatrix} 2 & 4 \\ 2 & -3 \\ 1 & -4 \end{pmatrix}, \mathbf{b} = \begin{pmatrix} 4 \\ 18 \\ 14 \end{pmatrix}$ . On a  $\text{rg}(\mathbb{A}) = \text{rg}([\mathbb{A}|\mathbf{b}]) = 2$  donc la solution est unique.
- 2.2.  $\mathbb{A} = \begin{pmatrix} 2 & -2 & 2 \\ -1 & 2 & 3 \\ 0 & -1 & -4 \\ -2 & 3 & 2 \end{pmatrix}, \mathbf{b} = \begin{pmatrix} 6 \\ 0 \\ -3 \\ -3 \end{pmatrix}$ . On a  $\text{rg}(\mathbb{A}) = \text{rg}([\mathbb{A}|\mathbf{b}]) = 2 < 3$  donc il y a une infinité de solutions.
- 2.3.  $\mathbb{A} = \begin{pmatrix} 2 & -2 & 2 \\ -1 & 2 & 3 \\ 0 & -1 & -4 \\ -2 & 3 & 2 \end{pmatrix}, \mathbf{b} = \begin{pmatrix} 6 \\ 0 \\ -4 \\ -3 \end{pmatrix}$ . On a  $\text{rg}(\mathbb{A}) = 2 \neq \text{rg}([\mathbb{A}|\mathbf{b}]) = 3$  donc il n'y a pas de solution.

3.  $m$  équations et  $n$  inconnues avec  $m < n$  :

- 3.1.  $\mathbb{A} = \begin{pmatrix} 2 & -1 & 2 \\ -1 & 2 & 2 \end{pmatrix}, \mathbf{b} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$ . On a  $\text{rg}(\mathbb{A}) = \text{rg}([\mathbb{A}|\mathbf{b}]) = 2 < 3$  donc il y a une infinité de solutions.
- 3.2.  $\mathbb{A} = \begin{pmatrix} 2 & -1 & 1 \\ 4 & -2 & 2 \end{pmatrix}, \mathbf{b} = \begin{pmatrix} 1 \\ 4 \end{pmatrix}$ . On a  $\text{rg}(\mathbb{A}) = 1 \neq \text{rg}([\mathbb{A}|\mathbf{b}]) = 2$  donc il n'y a pas de solution.

**Astuce**

Soit  $r$  le rang du système  $(S)$  et  $p$  le nombre d'inconnues.

- ★ Si  $r = p$ ,  $(S)$  a une unique solution,
- ★ si  $r < p$ ,  $(S)$  a une infinité de solutions. Les  $r$  inconnues qui figurent au début des  $r$  équations issues de la méthode du pivot de GAUSS sont les inconnues principales. Elles peuvent se calculer de façon unique en fonction des autres  $p - r$  inconnues.

*Le choix des inconnues principales d'un système est arbitraire, mais leur nombre est toujours le même.*

**EXEMPLE**

On cherche toutes les solutions du système linéaire homogène

$$(S) \quad \begin{cases} x_1 + x_2 + 3x_3 + x_4 = 0, \\ x_1 + 3x_2 + 2x_3 + 4x_4 = 0, \\ 2x_1 + x_3 - x_4 = 0. \end{cases}$$

Le système étant homogène, il est inutile d'écrire le terme source dans la méthode du pivot de GAUSS :

$$\mathbb{A} = \begin{pmatrix} 1 & 1 & 3 & 1 \\ 1 & 3 & 2 & 4 \\ 2 & 0 & 1 & -1 \end{pmatrix} \xrightarrow[L_2 \leftarrow L_2 - L_1]{L_3 \leftarrow L_3 - 2L_1} \begin{pmatrix} 1 & 1 & 3 & 1 \\ 0 & 2 & -1 & 3 \\ 0 & -2 & -5 & -3 \end{pmatrix} \xrightarrow[L_3 \leftarrow L_3 + L_2]{L_1 \leftarrow L_1 - L_2} \begin{pmatrix} 1 & 1 & 3 & 1 \\ 0 & 2 & -1 & 3 \\ 0 & 0 & -6 & 0 \end{pmatrix}$$

Le système admet une infinité de solutions de la forme  $(\frac{1}{2}\kappa, -\frac{3}{2}\kappa, 0, \kappa)$  avec  $\kappa \in \mathbb{R}$ .

**Astuce**

Pour résoudre un système  $(S)$  de  $m$  équations à  $n$  inconnues où  $m > n$  on considère un sous-système carré  $(S')$  de  $n$  équations à  $n$  inconnues et on résout ce système :

- ★ si  $(S')$  n'admet pas de solution, alors  $(S)$  non plus ;
- ★ si  $(S')$  admet une unique solution  $(c_1, c_2, \dots, c_n)$ , alors on vérifie si cette solution vérifie les autres  $m - n$  équations du système  $(S)$  :
  - ★ si oui, alors  $(S)$  admet l'unique solution  $(c_1, c_2, \dots, c_n)$ ,
  - ★ si non, alors  $(S)$  n'admet pas de solution ;
- ★ si  $(S')$  admet une infinité de solutions, on cherche parmi ces solutions celles qui vérifient également les autres équations de  $(S)$ .

**EXEMPLE**

Considérons le système de 4 équations à 3 inconnues

$$(S) \quad \begin{cases} x + y + z = 3, \\ x + 2y + 3z = 6, \\ -x - y + 2z = 0, \\ 3x + 2y - 4z = 1, \end{cases}$$

Pour résoudre  $(S)$ , on considère le sous-système carré d'ordre 3

$$(S') \quad \begin{cases} x + y + z = 3, \\ x + 2y + 3z = 6, \\ -x - y + 2z = 0, \end{cases}$$

qu'on peut résoudre par la méthode du pivot de GAUSS

$$\begin{cases} x + y + z = 3, \\ x + 2y + 3z = 6, \\ -x - y + 2z = 0, \end{cases} \xrightarrow[L_3 \leftarrow L_3 + L_1]{L_2 \leftarrow L_2 - L_1} \begin{cases} x + y + z = 3, \\ y + 2z = 3, \\ 3z = 3, \end{cases}$$

Ce sous-système admet l'unique solution  $(1, 1, 1)$ . On étudie alors si elle est aussi solution de l'équation de  $(S)$  qui n'apparaît pas dans  $(S')$  : pour  $(x, y, z) = (1, 1, 1)$  on a  $3x + 2y - 4z = 1$  donc le triplet  $(1, 1, 1)$  est solution de  $(S)$  et c'est l'unique.

## 5.3 Méthodes de résolution numériques

### ⚠ ATTENTION

Un système linéaire ne change pas de solution si on change l'ordre des équations. Cependant, l'ordre des équations peut changer totalement la solution donnée par une méthode numérique !

La solution du système  $\mathbb{A}\mathbf{x} = \mathbf{b}$  existe et est unique si et seulement si  $\mathbb{A}$  n'est pas singulière. En théorie, la solution peut être calculée en utilisant les formules de CRAMER. Si les  $n + 1$  déterminants sont calculés par le développement de LAPLACE, environ  $3(n + 1)!$  opérations sont nécessaires.<sup>1</sup>

Ce coût est trop élevé pour les grandes valeurs de  $n$  qu'on rencontre souvent en pratique. Deux classes de méthodes sont alors utilisées :

- ★ les méthodes directes, qui donnent la solution en un nombre fini d'étapes,
- ★ et les méthodes itératives, qui nécessitent (théoriquement) un nombre infini d'étapes.

Il faut être conscient que le choix entre méthodes directes et itératives dépend de nombreux critères : l'efficacité théorique de l'algorithme, le type de matrice, la capacité de stockage en mémoire, l'architecture de l'ordinateur. Notons enfin qu'un système associé à une matrice pleine ne peut pas être résolu par moins de  $n^2$  opérations. En effet, si les équations sont toutes couplées, on peut s'attendre à ce que chacun des  $n^2$  coefficients de la matrice soit impliqué au moins une fois dans une opération algébrique.

### 💡 Définition 5.11 (Conditionnement d'une matrice)

Le conditionnement d'une matrice  $\mathbb{A} \in \mathbb{R}^{n \times n}$  non singulière est défini par

$$K(\mathbb{A}) = \|\mathbb{A}\| \|\mathbb{A}^{-1}\| (\geq 1),$$

où  $\|\cdot\|$  est une norme matricielle subordonnée. En général,  $K(\mathbb{A})$  dépend du choix de la norme ; ceci est signalé en introduisant un indice dans la notation. Par exemple, on a les deux normes matricielles suivantes :

$$\|\mathbb{A}\|_1 = \max_{j=1,\dots,n} \sum_{i=1}^n |a_{ij}|, \quad \|\mathbb{A}\|_\infty = \max_{i=1,\dots,n} \sum_{j=1}^n |a_{ij}|.$$

### ✿ Remarque (Cas particulier)

Si  $\mathbb{A}$  est symétrique et définie positive<sup>2</sup>,

$$K_2(\mathbb{A}) = \|\mathbb{A}\|_2 \|\mathbb{A}^{-1}\|_2 = \frac{\lambda_{\max}}{\lambda_{\min}}$$

où  $\lambda_{\max}$  (resp.  $\lambda_{\min}$ ) est la plus grande (resp. petite) valeur propre de  $\mathbb{A}$ .

Considérons un système non singulier  $\mathbb{A}\mathbf{x} = \mathbf{b}$ . Si  $\delta\mathbf{b}$  est une perturbation de  $\mathbf{b}$  et si on résout  $\mathbb{A}\mathbf{y} = \mathbf{b} + \delta\mathbf{b}$ , on obtient par linéarité  $\mathbf{y} = \mathbf{x} + \delta\mathbf{x}$  avec  $\mathbb{A}\delta\mathbf{x} = \delta\mathbf{b}$ . La question est de savoir s'il est possible de majorer l'erreur relative  $\|\delta\mathbf{x}\|/\|\mathbf{x}\|$  sur la solution du système en fonction de l'erreur relative  $\|\delta\mathbf{b}\|/\|\mathbf{b}\|$  commise sur le second membre. Il est possible de démontrer que

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq K(\mathbb{A}) \frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}$$

où  $K(\mathbb{A})$  est le nombre de conditionnement de la matrice  $\mathbb{A}$ . On voit alors que plus le conditionnement de la matrice est grand, plus la solution du système linéaire est sensible aux perturbations des données. Cependant, le fait qu'un système linéaire soit bien conditionné n'implique pas nécessairement que sa solution soit calculée avec précision. Il faut en plus utiliser des algorithmes stables. Inversement, le fait d'avoir une matrice avec un grand conditionnement n'empêche pas nécessairement le système global d'être bien conditionné pour des choix particuliers du second membre.

Si  $\|\delta\mathbf{b}\|/\|\mathbf{b}\|$  est de l'ordre de la précision relative  $\eta = 10^{-p}$  du calculateur, alors  $\|\delta\mathbf{x}\|/\|\mathbf{x}\|$  pourrait, au pire, être égal à

$$K(\mathbb{A})\eta = 10^{\log_{10}(K(\mathbb{A}))} 10^{-p} = 10^{\log_{10}(K(\mathbb{A}) - p)}.$$

Si on calcule la solution du système linéaire avec un ordinateur à  $p$  chiffres significatifs en valeur décimale, on ne pourra pas garantir à priori plus de

$$E(p - \log_{10}(K(\mathbb{A})))$$

1. On entend par opération une somme, une soustraction, un produit ou une division.

2.  $\mathbb{A} \in \mathbb{R}^{n \times n}$  est

★ symétrique si  $a_{ij} = a_{ji}$  pour tout  $i, j = 1, \dots, n$ ,

★ définie positive si pour tout vecteur  $\mathbf{x} \in \mathbb{R}^n$  avec  $\mathbf{x} \neq \mathbf{0}$ ,  $\mathbf{x}^T \mathbb{A} \mathbf{x} > 0$ .

chiffres significatifs sur la solution. Si on applique cette règle au système linéaire de l'exemple, il est facile de vérifier que  $K(\mathbb{A}) \approx 10^7$ , par conséquent nous pouvons perdre jusqu'à 7 chiffres significatifs lors de sa résolution. Il faut donc un ordinateur calculant avec 10 chiffres significatifs pour être sûr d'obtenir les 3 premiers chiffres de la solution.

### EXEMPLE

Un exemple bien connu de matrice mal conditionnée est la matrice de HILBERT d'ordre  $n$  définie par  $a_{ij} = 1/(i+j-1)$  pour  $1 \leq i, j \leq n$ .

to do

## 5.3.1 Méthodes directes

### Méthode de Gauss

La méthode de GAUSS transforme le système  $\mathbb{A}\mathbf{x} = \mathbf{b}$  en un système équivalent (c'est-à-dire ayant la même solution) de la forme  $\mathbb{U}\mathbf{x} = \mathbf{y}$ , où  $\mathbb{U}$  est une matrice triangulaire supérieure et  $\mathbf{y}$  est un second membre convenablement modifié. Enfin on résout le système triangulaire  $\mathbb{U}\mathbf{x} = \mathbf{y}$ :  $x_n = \frac{y_n}{a_{nn}}$  et on déduit les inconnues  $x_{n-1}, x_{n-2}, \dots, x_1$  grâce à la relation

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

### Définition 5.12 (Méthode du pivot de GAUSS)

Soit  $\mathbb{A} = (a_{ij})_{\substack{1 \leq i \leq n \\ 1 \leq j \leq p}}$  la matrice des coefficients du système  $\mathbb{A}\mathbf{x} = \mathbf{b}$ .

*Étape k* : en permutant éventuellement deux lignes du système, on peut supposer  $a_{kk} \neq 0$  (appelé pivot de l'étape  $k$ ). On transforme toutes les lignes  $L_i$  avec  $i > k$  comme suit :

$$L_i \leftarrow L_i - \frac{a_{ik}}{a_{kk}} L_k.$$

En réitérant le procédé pour  $k$  de 1 à  $n-1$ , on aboutit à un système triangulaire supérieur.

### EXEMPLE

Résolution du système linéaire

$$\begin{cases} x_1 + 2x_2 + 3x_3 + 4x_4 = 1, \\ 2x_1 + 3x_2 + 4x_3 + x_4 = 2, \\ 3x_1 + 4x_2 + x_3 + 2x_4 = 3, \\ 4x_1 + x_2 + 2x_3 + 3x_4 = 4. \end{cases}$$

$$[\mathbb{A}|\mathbf{b}] = \left( \begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 2 & 3 & 4 & 1 & 2 \\ 3 & 4 & 1 & 2 & 3 \\ 4 & 1 & 2 & 3 & 4 \end{array} \right) \xrightarrow{\substack{L_2 \leftarrow L_2 - 2L_1 \\ L_3 \leftarrow L_3 - 3L_1 \\ L_4 \leftarrow L_4 - 4L_1}} \left( \begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 0 & -1 & -2 & -7 & 0 \\ 0 & -2 & -8 & -10 & 0 \\ 0 & -7 & -10 & -13 & 0 \end{array} \right) \xrightarrow{\substack{L_3 \leftarrow L_3 - 2L_2 \\ L_4 \leftarrow L_4 - 7L_2}} \left( \begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 0 & -1 & -2 & -7 & 0 \\ 0 & 0 & -4 & 4 & 0 \\ 0 & 0 & 4 & 36 & 0 \end{array} \right) \xrightarrow{L_4 \leftarrow L_4 + L_3} \left( \begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 0 & -1 & -2 & -7 & 0 \\ 0 & 0 & -4 & 4 & 0 \\ 0 & 0 & 0 & 40 & 0 \end{array} \right)$$

donc  $x_4 = 0$ ,  $x_3 = 0$ ,  $x_2 = 0$  et  $x_1 = 1$ .

Si on a plusieurs systèmes dont seul le second membre change, il peut être utile de factoriser une fois pour toute la matrice  $\mathbb{A}$  et résoudre ensuite des systèmes triangulaires.

### Factorisation $\mathbb{LU}$ et systèmes linéaires

Soit  $\mathbb{A} \in \mathbb{R}^{n \times n}$ . Supposons qu'il existe deux matrices,  $\mathbb{L}$  et  $\mathbb{U}$ , respectivement triangulaire inférieure et supérieure, telles que  $\mathbb{A} = \mathbb{LU}$ . On appelle cela une factorisation (*ou décomposition*)  $\mathbb{LU}$  de  $\mathbb{A}$ . Si  $\mathbb{A}$  est régulière (*i.e.* non singulière), alors  $\mathbb{L}$  et  $\mathbb{U}$  le sont aussi, et leurs termes diagonaux sont donc non nuls. Dans ce cas, résoudre  $\mathbb{A}\mathbf{x} = \mathbf{b}$  revient à résoudre deux systèmes triangulaires : d'abord le système  $\mathbb{L}\mathbf{y} = \mathbf{b}$ , puis  $\mathbb{U}\mathbf{x} = \mathbf{y}$ .

Les deux systèmes sont faciles à résoudre :

- \* **système triangulaire inférieur**  $\mathbb{L}\mathbf{y} = \mathbf{b}$  :  $\mathbb{L}$  étant triangulaire inférieure, la première ligne du système  $\mathbb{L}\mathbf{y} = \mathbf{b}$  est de la forme

$$l_{11}y_1 = b_1$$

ce qui donne la valeur de  $y_1$  puisque  $l_{11} \neq 0$ . En substituant cette valeur de  $y_1$  dans les  $n - 1$  équations suivantes, on obtient un nouveau système dont les inconnues sont  $y_2, \dots, y_n$ , pour lesquelles on peut faire de même. En procédant équation par équation, on calcule ainsi toutes les inconnues par l'algorithme dit *de descente* :

$$\begin{aligned} y_1 &= \frac{1}{l_{11}}b_1, \\ y_i &= \frac{1}{l_{ii}} \left( b_i - \sum_{j=1}^{i-1} l_{ij}y_j \right), \quad i = 2, \dots, n \end{aligned}$$

Évaluons le nombre d'opérations requis : on effectue  $i - 1$  sommes,  $i - 1$  produits et 1 division pour calculer l'inconnue  $y_i$ . Le nombre total d'opérations est donc

$$\sum_{i=1}^n 1 + 2 \sum_{i=1}^n (i - 1) = n^2.$$

- \* **système triangulaire supérieur**  $\mathbb{U}\mathbf{x} = \mathbf{y}$  : on peut résoudre le système  $\mathbb{U}\mathbf{x} = \mathbf{y}$  de manière similaire. Cette fois, on commence par déterminer  $x_n$  puis, de proche en proche, les autres inconnues  $x_i$  de  $i = n - 1$  à  $i = 1$ . En procédant équation par équation, on calcule ainsi toutes les inconnues par l'algorithme dit *de remontée* :

$$\begin{aligned} x_n &= \frac{1}{u_{nn}}b_1, \\ x_i &= \frac{1}{u_{ii}} \left( b_i - \sum_{j=i+1}^n u_{ij}x_j \right), \quad i = n - 1, \dots, 1 \end{aligned}$$

Il nécessite également  $n^2$  opérations.

Il reste à présent à trouver un algorithme qui permette le calcul effectif des facteurs  $\mathbb{L}$  et  $\mathbb{U}$ .

- \* **factorisation  $\mathbb{LU}$**  : en fixant la valeur 1 pour les  $n$  éléments diagonaux de  $\mathbb{L}$ , la matrice  $\mathbb{U}$  peut être déterminée avec l'algorithme de Gauss et la matrice  $\mathbb{L}$  contient les coefficients multiplicateurs de chaque ligne  $i$  à l'étape  $k$ . Les termes non nuls de  $\mathbb{U}$  et les termes non nuls en-dessous de la diagonale principale de  $\mathbb{L}$  sont mémorisés encore dans la matrice  $\mathbb{A}$  et sont ainsi calculées :

posons  $\mathbb{A}^{(1)} = \mathbb{A}$  i.e.  $a_{ij} = a_{ij}$  pour  $i, j = 1, \dots, n$ ;

**for**  $k = 1$  à  $n - 1$  **do**

**for**  $i = k + 1$  à  $n$  **do**

$$a_{ik}^{(k+1)} \leftarrow \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}$$

{Il s'agit de  $\ell_{ik}$  mémorisé dans  $a_{ik}^{(k+1)}$ }

**for**  $j = k + 1$  à  $n$  **do**

$$a_{ij}^{(k+1)} \leftarrow a_{ij}^{(k)} - a_{ik}^{(k+1)}a_{kj}^{(k)}$$

**end for**

**end for**

**end for**

Attention, à chaque étape  $k$ , le terme  $a_{kk}^{(k)}$ , appelé pivot, doit être non nul!



### Proposition 5.13

Pour une matrice quelconque  $\mathbb{A} \in \mathbb{R}^{n \times n}$ , la factorisation  $\mathbb{LU}$  existe et est unique si et seulement si les sous-matrices principales  $\mathbb{A}_i$  de  $\mathbb{A}$  d'ordre  $i = 1, \dots, n - 1$  (celles que l'on obtient en restreignant  $\mathbb{A}$  à ses  $i$  premières lignes et colonnes) ne sont pas singulières (autrement dit si les mineurs principaux, i.e. les déterminants des sous-matrices principales, sont non nuls).

On peut identifier des classes de matrices particulières pour lesquelles les hypothèses de cette proposition sont satisfaites :



### Proposition 5.14

Si la matrice  $\mathbb{A} \in \mathbb{R}^{n \times n}$  est symétrique et définie positive ou si est à diagonale dominante<sup>3</sup> alors la factorisation  $\mathbb{LU}$  existe et est unique.

---

3.  $\mathbb{A} \in \mathbb{R}^{n \times n}$  est

- \* symétrique si  $a_{ij} = a_{ji}$  pour tout  $i, j = 1, \dots, n$ ,
- \* définie positive si pour tout vecteurs  $\mathbf{x} \in \mathbb{R}^n$  avec  $\mathbf{x} \neq \mathbf{0}$ ,  $\mathbf{x}^T \mathbb{A} \mathbf{x} > 0$ ,

Une technique qui permet d'effectuer la factorisation  $\mathbb{L}\mathbb{U}$  pour toute matrice  $\mathbb{A}$  inversible, même quand les hypothèses de cette proposition ne sont pas vérifiées, est la méthode du pivot par ligne : il suffit d'effectuer une permutation convenable des lignes de la matrice originale  $\mathbb{A}$  à chaque étape  $k$  où un terme diagonal  $a_{kk}$  s'annule.

### Définition 5.15 (Algorithme de GAUSS avec pivot)

Dans la méthode d'élimination de GAUSS les pivot  $a_{kk}^{(k)}$  doivent être différents de zéro. Si la matrice est inversible mais un pivot est zéro (ou numériquement proche de zéro), on peut permute deux lignes avant de poursuivre la factorisation. Concrètement, à chaque étape on cherche à avoir le pivot de valeur absolue la plus grande possible. L'algorithme modifié s'écrit alors

```

for  $k = 1$  to  $n - 1$  do
  for  $i = k + 1$  to  $n$  do
    Chercher  $\bar{r}$  tel que  $|a_{\bar{r}k}^{(k)}| = \max_{r=k,\dots,n} |a_{rk}^{(k)}|$  et échanger la ligne  $k$  avec la ligne  $\bar{r}$ 
     $\ell_{ik} \leftarrow \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}$ 
    for  $j = k + 1$  to  $n$  do
       $a_{ij}^{(k+1)} \leftarrow a_{ij}^{(k)} - \ell_{ik} a_{kj}^{(k)}$ 
    end for
  end for
end for

```

Une fois calculées les matrices  $\mathbb{L}$  et  $\mathbb{U}$  et la matrice des permutations  $\mathbb{P}$  (*i.e.* la matrice telle que  $\mathbb{PA} = \mathbb{LU}$ ), résoudre le système linéaire consiste simplement à résoudre successivement le système triangulaire inférieur  $\mathbb{Ly} = \mathbb{Pb}$  puis le système triangulaire supérieure  $\mathbb{Ux} = \mathbb{y}$ .

TO DO : construire la matrice  $\mathbb{P}$  des pivots

### Propriété 5.16 (Déterminant)

La factorisation  $\mathbb{LU}$  permet de calculer le déterminant de  $\mathbb{A}$  en  $O(n^3)$  car  $\det(\mathbb{A}) = \det(\mathbb{L}) \det(\mathbb{U}) = \prod_{k=1}^n u_{kk}$ .

### Propriété 5.17 (Inverse d'une matrice)

Le calcul explicite de l'inverse d'une matrice peut être effectué en utilisant la factorisation  $\mathbb{LU}$  comme suit. En notant  $\mathbb{X}$  l'inverse d'une matrice régulière  $\mathbb{A} \in \mathbb{R}^{n \times n}$ , les vecteurs colonnes de  $\mathbb{X}$  sont les solutions des systèmes linéaires

$$\mathbb{Ax}_i = \mathbf{e}_i, \quad \text{pour } i = 1, \dots, n.$$

En supposant que  $\mathbb{PA} = \mathbb{LU}$ , où  $\mathbb{P}$  est la matrice de changement de pivot partiel, on doit résoudre  $2n$  systèmes triangulaires de la forme

$$\mathbb{Ly}_i = \mathbb{Pe}_i, \quad \mathbb{Ux}_i = \mathbb{y}_i, \quad \text{pour } i = 1, \dots, n.$$

c'est-à-dire une suite de systèmes linéaires ayant la même matrice mais des seconds membres différents.

### EXEMPLE

Soit les systèmes linéaires

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \\ 3 & 4 & 1 & 2 \\ 4 & 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} \quad \text{et} \quad \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \\ 3 & 4 & 1 & 2 \\ 4 & 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 10 \\ 10 \\ 10 \\ 10 \end{pmatrix}.$$

1. Résoudre les systèmes linéaires par la méthode du pivot de GAUSS.
2. Factoriser la matrice  $\mathbb{A}$  (sans utiliser la technique du pivot) et résoudre les systèmes linéaires.
3. Calculer le déterminant de  $\mathbb{A}$ .
4. Calculer  $\mathbb{A}^{-1}$ .

\* à diagonale dominante par lignes si  $|a_{ii}| \geq \sum_{j \neq i}^n |a_{ij}|$ , pour  $i = 1, \dots, n$  (à diagonale dominante stricte par lignes si l'inégalité est stricte),

\* à diagonale dominante par colonnes si  $|a_{ii}| \geq \sum_{j \neq i}^n |a_{ji}|$ , pour  $i = 1, \dots, n$  (à diagonale dominante stricte par colonnes si l'inégalité est stricte),

## 1. Résolution par la méthode du pivot de GAUSS du premier système

$$[\mathbb{A}|\mathbf{b}] = \left( \begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 2 & 3 & 4 & 1 & 2 \\ 3 & 4 & 1 & 2 & 3 \\ 4 & 1 & 2 & 3 & 4 \end{array} \right) \xrightarrow{\substack{L_2 \leftarrow L_2 - 2L_1 \\ L_3 \leftarrow L_3 - 3L_1 \\ L_4 \leftarrow L_4 - 4L_1}} \left( \begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 0 & -1 & -2 & -7 & 0 \\ 0 & -2 & -8 & -10 & 0 \\ 0 & -7 & -10 & -13 & 0 \end{array} \right) \xrightarrow{\substack{L_3 \leftarrow L_3 - 2L_2 \\ L_4 \leftarrow L_4 - 7L_2}} \left( \begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 0 & -1 & -2 & -7 & 0 \\ 0 & 0 & -4 & 4 & 0 \\ 0 & 0 & 0 & 40 & 0 \end{array} \right) \xrightarrow{L_4 \leftarrow L_4 + L_3} \left( \begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 0 & -1 & -2 & -7 & 0 \\ 0 & 0 & -4 & 4 & 0 \\ 0 & 0 & 0 & 40 & 0 \end{array} \right)$$

donc

$$x_4 = 0, \quad x_3 = 0, \quad x_2 = 0, \quad x_1 = 1.$$

## Résolution par la méthode du pivot de GAUSS du second système

$$(\mathbb{A}|\mathbf{b}) = \left( \begin{array}{cccc|c} 1 & 2 & 3 & 4 & 10 \\ 2 & 3 & 4 & 1 & 10 \\ 3 & 4 & 1 & 2 & 10 \\ 4 & 1 & 2 & 3 & 10 \end{array} \right) \xrightarrow{\substack{L_2 \leftarrow L_2 - 2L_1 \\ L_3 \leftarrow L_3 - 3L_1 \\ L_4 \leftarrow L_4 - 4L_1}} \left( \begin{array}{cccc|c} 1 & 2 & 3 & 4 & 10 \\ 0 & -1 & -2 & -7 & -10 \\ 0 & -2 & -8 & -10 & -20 \\ 0 & -7 & -10 & -13 & -30 \end{array} \right) \xrightarrow{\substack{L_3 \leftarrow L_3 - 2L_2 \\ L_4 \leftarrow L_4 - 7L_2}} \left( \begin{array}{cccc|c} 1 & 2 & 3 & 4 & 10 \\ 0 & -1 & -2 & -7 & -10 \\ 0 & 0 & -4 & 4 & 0 \\ 0 & 0 & 0 & 40 & 40 \end{array} \right) \xrightarrow{L_4 \leftarrow L_4 + L_3} \left( \begin{array}{cccc|c} 1 & 2 & 3 & 4 & 10 \\ 0 & -1 & -2 & -7 & -10 \\ 0 & 0 & -4 & 4 & 0 \\ 0 & 0 & 0 & 40 & 40 \end{array} \right)$$

donc

$$\begin{cases} x_1 + 2x_2 + 3x_3 + 4x_4 = 10 \\ -x_2 - 2x_3 - 7x_4 = -10 \\ -4x_3 + 4x_4 = 0 \\ 40x_4 = 40 \end{cases} \implies x_4 = 1, \quad x_3 = 1, \quad x_2 = 1, \quad x_1 = 1.$$

2. Factorisation de la matrice  $\mathbb{A}$ :

$$\left( \begin{array}{cccc} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \\ 3 & 4 & 1 & 2 \\ 4 & 1 & 2 & 3 \end{array} \right) \xrightarrow{\substack{L_2 \leftarrow L_2 - 2L_1 \\ L_3 \leftarrow L_3 - 3L_1 \\ L_4 \leftarrow L_4 - 4L_1}} \left( \begin{array}{cccc} 1 & 2 & 3 & 4 \\ 2 & -1 & -2 & -7 \\ 3 & -2 & -8 & -10 \\ 4 & -7 & -10 & -13 \end{array} \right) \xrightarrow{\substack{L_3 \leftarrow L_3 - 2L_2 \\ L_4 \leftarrow L_4 - 7L_2}} \left( \begin{array}{cccc} 1 & 2 & 3 & 4 \\ 2 & -1 & -2 & -7 \\ 3 & 2 & -4 & 4 \\ 4 & 7 & 4 & 36 \end{array} \right) \xrightarrow{L_4 \leftarrow L_4 + L_3} \left( \begin{array}{cccc} 1 & 2 & 3 & 4 \\ 2 & -1 & -2 & -7 \\ 3 & 2 & -4 & 4 \\ 4 & 7 & -1 & 40 \end{array} \right)$$

donc

$$\mathbb{L} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 2 & 1 & 0 \\ 4 & 7 & -1 & 1 \end{pmatrix} \quad \mathbb{U} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & -1 & -2 & -7 \\ 0 & 0 & -4 & 4 \\ 0 & 0 & 0 & 40 \end{pmatrix}$$

Pour résoudre le premier système linéaire on résout les systèmes triangulaires  $\mathbb{L}\mathbf{y} = \mathbf{b}$ 

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 2 & 1 & 0 \\ 4 & 7 & -1 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} \implies y_1 = 1, \quad y_2 = 0, \quad y_3 = 0, \quad y_4 = 0$$

et  $\mathbb{U}\mathbf{x} = \mathbf{y}$ 

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & -1 & -2 & -7 \\ 0 & 0 & -4 & 4 \\ 0 & 0 & 0 & 40 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \implies x_4 = 0, \quad x_3 = 0, \quad x_2 = 0, \quad x_1 = 1.$$

Pour résoudre le second système linéaire on résout les systèmes triangulaires  $\mathbb{L}\mathbf{y} = \mathbf{b}$

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 2 & 1 & 0 \\ 4 & 7 & -1 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 10 \\ 10 \\ 10 \\ 10 \end{pmatrix} \implies y_1 = 10, \quad y_2 = -10, \quad y_3 = 0, \quad y_4 = 40$$

et  $\mathbb{U}\mathbf{x} = \mathbf{y}$

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & -1 & -2 & -7 \\ 0 & 0 & -4 & 4 \\ 0 & 0 & 0 & 40 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 10 \\ -10 \\ 0 \\ 40 \end{pmatrix} \implies x_4 = 1, \quad x_3 = 1, \quad x_2 = 1, \quad x_1 = 1.$$

3. Le déterminant de  $\mathbb{A}$  est  $u_{11}u_{22}u_{33}u_{44} = 1 \times (-1) \times (-4) \times 40 = 160$ .

4. Pour calculer  $\mathbb{A}^{-1}$  on résout les quatre systèmes linéaires

$$\begin{array}{l} \left( \begin{array}{cccc} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \\ 3 & 4 & 1 & 2 \\ 4 & 1 & 2 & 3 \end{array} \right) \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \text{ i.e. } \left( \begin{array}{cccc} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 2 & 1 & 0 \\ 4 & 7 & -1 & 1 \end{array} \right) \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \implies \begin{pmatrix} 1 \\ -2 \\ 1 \\ 11 \end{pmatrix} \text{ puis } \left( \begin{array}{cccc} 1 & 2 & 3 & 4 \\ 0 & -1 & -2 & -7 \\ 0 & 0 & -4 & 4 \\ 0 & 0 & 0 & 40 \end{array} \right) \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ -2 \\ 1 \\ 11 \end{pmatrix} \implies \begin{pmatrix} -9/40 \\ 1/40 \\ 1/40 \\ 11/40 \end{pmatrix} \\ \left( \begin{array}{cccc} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \\ 3 & 4 & 1 & 2 \\ 4 & 1 & 2 & 3 \end{array} \right) \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \text{ i.e. } \left( \begin{array}{cccc} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 2 & 1 & 0 \\ 4 & 7 & -1 & 1 \end{array} \right) \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \implies \begin{pmatrix} 0 \\ 1 \\ -2 \\ -9 \end{pmatrix} \text{ puis } \left( \begin{array}{cccc} 1 & 2 & 3 & 4 \\ 0 & -1 & -2 & -7 \\ 0 & 0 & -4 & 4 \\ 0 & 0 & 0 & 40 \end{array} \right) \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ -2 \\ -9 \end{pmatrix} \implies \begin{pmatrix} 1/40 \\ 1/40 \\ 11/40 \\ -9/40 \end{pmatrix} \\ \left( \begin{array}{cccc} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \\ 3 & 4 & 1 & 2 \\ 4 & 1 & 2 & 3 \end{array} \right) \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \text{ i.e. } \left( \begin{array}{cccc} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 2 & 1 & 0 \\ 4 & 7 & -1 & 1 \end{array} \right) \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \implies \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} \text{ puis } \left( \begin{array}{cccc} 1 & 2 & 3 & 4 \\ 0 & -1 & -2 & -7 \\ 0 & 0 & -4 & 4 \\ 0 & 0 & 0 & 40 \end{array} \right) \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} \implies \begin{pmatrix} 1/40 \\ 11/40 \\ -9/40 \\ 1/40 \end{pmatrix} \\ \left( \begin{array}{cccc} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \\ 3 & 4 & 1 & 2 \\ 4 & 1 & 2 & 3 \end{array} \right) \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \text{ i.e. } \left( \begin{array}{cccc} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 2 & 1 & 0 \\ 4 & 7 & -1 & 1 \end{array} \right) \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \implies \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \text{ puis } \left( \begin{array}{cccc} 1 & 2 & 3 & 4 \\ 0 & -1 & -2 & -7 \\ 0 & 0 & -4 & 4 \\ 0 & 0 & 0 & 40 \end{array} \right) \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \implies \begin{pmatrix} 11/40 \\ -9/40 \\ 1/40 \\ 1/40 \end{pmatrix} \end{array}$$

et finalement

$$\mathbb{A}^{-1} = \begin{pmatrix} -9/40 & 1/40 & 1/40 & 11/40 \\ 1/40 & 1/40 & 11/40 & -9/40 \\ 1/40 & 11/40 & -9/40 & 1/40 \\ 11/40 & -9/40 & 1/40 & 1/40 \end{pmatrix} = \frac{1}{40} \begin{pmatrix} -9 & 1 & 1 & 11 \\ 1 & 1 & 11 & -9 \\ 11 & 11 & -9 & 1 \\ 11 & -9 & 1 & 1 \end{pmatrix}.$$

### 5.3.2 Méthodes itératives

On n'a décrit qu'un seul algorithme de résolution, l'algorithme de GAUSS. Or cet algorithme est bien insuffisant pour résoudre numériquement, c'est-à-dire sur ordinateur, les énormes systèmes linéaires rencontrés dans la pratique. L'analyse numérique matricielle est l'étude d'algorithmes efficaces dans le but de résoudre effectivement et efficacement de tels systèmes. C'est un vaste champ de recherche toujours très actif de nos jours.

Une méthode itérative pour le calcul de la solution d'un système linéaire  $\mathbb{A}\mathbf{x} = \mathbf{b}$  avec  $\mathbb{A} \in \mathbb{R}^{n \times n}$  est une méthode qui construit une suite de vecteurs  $\mathbf{x}^{(k)} = (x_1^{(k)}, x_2^{(k)}, \dots, x_n^{(k)})^T \in \mathbb{R}^n$  convergent vers le vecteur solution exacte  $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$  pour tout vecteur initiale  $\mathbf{x}^{(0)} = (x_1^{(0)}, x_2^{(0)}, \dots, x_n^{(0)})^T \in \mathbb{R}^n$  lorsque  $k$  tend vers  $+\infty$ . Dans ces notes on ne verra que deux méthodes itératives :

- \* la méthode de JACOBI,
- \* la méthode de GAUSS-SEIDEL.

#### Définition 5.18 (Méthode de JACOBI)

Soit  $\mathbf{x}^0 = (x_1^0, x_2^0, \dots, x_n^0)$  un vecteur donné. La méthode de JACOBI définit la composante  $x_i^{k+1}$  du vecteur  $\mathbf{x}^{k+1}$  à partir des composantes  $x_j^k$  du vecteur  $\mathbf{x}^k$  pour  $j \neq i$  de la manière suivante :

$$x_i^{k+1} = \frac{b_i - \sum_{j=1, j \neq i}^n a_{ij} x_j^k}{a_{ii}}, \quad i = 1, \dots, n$$

$$\mathbf{x}^{(k)} = \begin{pmatrix} x_1^{(k)} \\ x_2^{(k)} \\ \vdots \\ x_{i-1}^{(k)} \\ x_i^{(k)} \\ x_{i+1}^{(k)} \\ \vdots \\ x_n^{(k)} \end{pmatrix} \quad \mathbf{x}^{(k+1)} = \begin{pmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \\ \vdots \\ x_{i-1}^{(k+1)} \\ x_i^{(k+1)} \\ x_{i+1}^{(k+1)} \\ \vdots \\ x_n^{(k+1)} \end{pmatrix}$$

### Proposition 5.19

Si la matrice  $\mathbb{A}$  est à diagonale dominante stricte, la méthode de JACOBI converge.

La méthode de GAUSS-SIDEL est une amélioration de la méthode de JACOBI dans laquelle les valeurs calculées sont utilisées au fur et à mesure du calcul et non à l'issue d'une itération comme dans la méthode de JACOBI.

### Définition 5.20 (Méthode de GAUSS-SIDEL)

Soit  $\mathbf{x}^0 = (x_1^0, x_2^0, \dots, x_n^0)$  un vecteur donné. La méthode de GAUSS-SIDEL définit la composante  $x_i^{k+1}$  du vecteur  $\mathbf{x}^{k+1}$  à partir des composantes  $x_j^{k+1}$  du vecteur  $\mathbf{x}^{k+1}$  pour  $j < i$  et des composantes  $x_j^k$  du vecteur  $\mathbf{x}^k$  pour  $j \geq i$  de la manière suivante :

$$x_i^{k+1} = \frac{b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{k+1} - \sum_{j=i+1}^n a_{ij} x_j^k}{a_{ii}}, \quad i = 1, \dots, n$$

$$\mathbf{x}^{(k)} = \begin{pmatrix} x_1^{(k)} \\ x_2^{(k)} \\ \vdots \\ x_{i-1}^{(k)} \\ x_i^{(k)} \\ x_{i+1}^{(k)} \\ \vdots \\ x_n^{(k)} \end{pmatrix} \quad \mathbf{x}^{(k+1)} = \begin{pmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \\ \vdots \\ x_{i-1}^{(k+1)} \\ x_i^{(k+1)} \\ x_{i+1}^{(k+1)} \\ \vdots \\ x_n^{(k+1)} \end{pmatrix}$$

### Proposition 5.21

Si la matrice  $\mathbb{A}$  est à diagonale dominante stricte ou si elle est symétrique et définie positive, la méthode de GAUSS-SEIDEL converge.

Il n'y a pas de résultat général établissant que la méthode de GAUSS-SEIDEL converge toujours plus vite que celle de JACOBI. On peut cependant l'affirmer dans certains cas, comme le montre la proposition suivante

### Proposition 5.22

Soit  $\mathbb{A}$  une matrice tridiagonale de taille  $n \times n$  inversible dont les coefficients diagonaux sont tous non nuls. Alors les méthodes de JACOBI et de GAUSS-SEIDEL sont soit toutes les deux convergentes soit toutes les deux divergentes. En cas de convergence, la méthode de GAUSS-SEIDEL est plus rapide que celle de JACOBI.

### EXEMPLE

Considérons le système linéaire

$$\begin{pmatrix} 4 & 2 & 1 \\ -1 & 2 & 0 \\ 2 & 1 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \\ 9 \end{pmatrix}$$

mis sous la forme

$$\begin{cases} x = 1 - \frac{y}{2} - \frac{z}{4}, \\ y = 1 + \frac{x}{2}, \\ z = \frac{9}{4} - \frac{x}{2} - \frac{y}{4}. \end{cases}$$

Soit  $\mathbf{x}^{(0)} = (0, 0, 0)$  le vecteur initial.

- ★ En calculant les itérées avec la méthode de JACOBI on trouve

$$\begin{aligned} \mathbf{x}^{(1)} &= \begin{pmatrix} 1 - \frac{0}{2} - \frac{0}{4} \\ 1 + \frac{0}{2} \\ \frac{9}{4} - \frac{0}{2} - \frac{0}{4} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \frac{9}{4} \end{pmatrix}, & \mathbf{x}^{(2)} &= \begin{pmatrix} 1 - \frac{1}{2} - \frac{\frac{9}{4}}{4} \\ 1 + \frac{1}{2} \\ \frac{9}{4} - \frac{1}{2} - \frac{1}{4} \end{pmatrix} = \begin{pmatrix} -\frac{1}{16} \\ \frac{3}{2} \\ \frac{3}{2} \end{pmatrix}, \\ \mathbf{x}^{(3)} &= \begin{pmatrix} 1 - \frac{\frac{3}{2}}{2} - \frac{\frac{3}{2}}{4} \\ 1 + \frac{-\frac{1}{16}}{2} \\ \frac{9}{4} - \frac{-\frac{1}{16}}{2} - \frac{\frac{3}{2}}{4} \end{pmatrix} = \begin{pmatrix} -\frac{1}{8} \\ -\frac{1}{32} \\ \frac{61}{32} \end{pmatrix}, & \mathbf{x}^{(4)} &= \begin{pmatrix} 1 - \frac{-\frac{1}{32}}{2} - \frac{\frac{61}{32}}{4} \\ 1 + \frac{-\frac{1}{8}}{2} \\ \frac{9}{4} - \frac{-\frac{1}{8}}{2} - \frac{-\frac{1}{32}}{4} \end{pmatrix} = \begin{pmatrix} \frac{5}{128} \\ \frac{15}{16} \\ \frac{265}{128} \end{pmatrix}. \end{aligned}$$

La suite  $\mathbf{x}^{(k)}$  converge vers  $(0, 1, 2)$  la solution du système.

- ★ En calculant les itérées avec la méthode de GAUSS-SEIDEL on trouve

$$\mathbf{x}^{(1)} = \begin{pmatrix} 1 - \frac{0}{2} - \frac{0}{4} \\ 1 + \frac{1}{2} \\ \frac{9}{4} - \frac{1}{2} - \frac{\frac{3}{2}}{4} \end{pmatrix} = \begin{pmatrix} 1 \\ \frac{3}{2} \\ \frac{11}{8} \end{pmatrix}, \quad \mathbf{x}^{(2)} = \begin{pmatrix} 1 - \frac{\frac{3}{2}}{2} - \frac{\frac{11}{8}}{4} \\ 1 + \frac{-\frac{3}{2}}{2} \\ \frac{9}{4} - \frac{-\frac{3}{2}}{2} - \frac{\frac{61}{64}}{4} \end{pmatrix} = \begin{pmatrix} -\frac{3}{32} \\ \frac{61}{64} \\ \frac{527}{256} \end{pmatrix}, \quad \mathbf{x}^{(3)} = \begin{pmatrix} 1 - \frac{-\frac{3}{32}}{2} - \frac{\frac{61}{64}}{4} \\ 1 + \frac{\frac{9}{1024}}{2} \\ \frac{9}{4} - \frac{\frac{9}{1024}}{2} - \frac{\frac{2047}{2048}}{4} \end{pmatrix} = \begin{pmatrix} \frac{9}{1024} \\ \frac{2047}{2048} \\ \frac{16349}{8192} \end{pmatrix},$$

La suite  $\mathbf{x}^{(k)}$  converge vers  $(0, 1, 2)$  la solution du système.





## Exercices



### Utilisation de fonctions prédefinies

#### ★ Exercice 5.1 (Système linéaire, existence et unicité)

Considérons le système linéaire de 3 équations en les 3 inconnues  $x_1, x_2, x_3$  suivant :

$$\begin{cases} x_1 - x_2 + x_3 = 0 \\ 10x_2 + 25x_3 = 90 \\ 20x_1 + 10x_2 = 80. \end{cases}$$

Pour résoudre le système linéaire on commence par définir la matrice  $\mathbb{A}$  des coefficients du système et le vecteur colonne  $\mathbf{b}$  contenant le terme source.

Méthode 1. On calcule la matrice inverse  $\mathbb{A}^{-1}$  et on pose  $\mathbf{x} = \mathbb{A}^{-1}\mathbf{b}$  (méthode déconseillée).

Méthode 2. On utilise l'opérateur *backslash*.

Méthode 3. On définit la matrice augmentée  $[\mathbb{A}|\mathbf{b}]$  et on applique la méthode de GAUSS-JORDAN pour obtenir la forme échelonnée (instruction `rref(Aaug)`).

Dans tous les cas, on teste la solution obtenue en calculant  $\|\mathbb{A}\mathbf{x} - \mathbf{b}\|_2$ .

#### Correction

```
A = [ 1 -1 1; 0 10 25; 20 10 0]
b = [0; 90; 80]

% Methode 1
x = inv(A)*b
norm(A*x-b)
% while mathematically correct, computing the inverse of a matrix is
% computationally inefficient, and not recommended most of the time.

% Methode 2
x = A\b
norm(A*x-b)

% Methode 3
Aaug=[A b]
RRAaug=rref(Aaug)
x=RRAaug(:,4)
norm(A*x-b)
```

#### ★ Exercice 5.2 (Système linéaire, non existence)

Considérons le système linéaire de 3 équations en les 3 inconnues  $x_1, x_2, x_3$  suivant :

$$\begin{cases} 3x_1 + 2x_2 + x_3 = 3 \\ 2x_1 + x_2 + x_3 = 0 \\ 6x_1 + 2x_2 + 4x_3 = 6. \end{cases}$$

Pour résoudre le système linéaire on commence par définir la matrice  $\mathbb{A}$  des coefficients du système et le vecteur colonne  $\mathbf{b}$  contenant le terme source.

1. On définit la matrice augmentée  $[\mathbb{A}|\mathbf{b}]$  et on applique la méthode de GAUSS-JORDAN pour obtenir la forme échelonnée (instruction `rref(Aaug)`). Pourquoi peut-on conclure que le système n'a pas de solution ?
2. Octave nous donne malgré tout une solution ! Vérifiez-le avec l'opérateur *backslash*.
3. Que se passe-t-il si on essaye de calculer la matrice inverse  $\mathbb{A}^{-1}$  et poser ensuite  $\mathbf{x} = \mathbb{A}^{-1}\mathbf{b}$  ?

Dans tous les cas, on teste la solution obtenue en calculant  $\|\mathbb{A}\mathbf{x} - \mathbf{b}\|_2$ .

**Correction**

```

A = [ 3 2 1; 2 1 1; 6 2 4]
b = [3; 0; 6]

% Point 1
rref([A ,b])
% the last line of this matrix states that 0 = 1. That is not true, which
% means there is no solution.

% Point 2
x = A\b
norm(A*x-b)

% Point 3
invA=inv(A)
x = invA*b
norm(A*x-b)

```

**Méthode de Gauss pour des systèmes carrés****Exercice 5.3**

Résoudre les systèmes linéaires suivants :

$$\textcircled{1} \begin{cases} x_1 + 2x_2 - x_3 = 2 \\ x_1 - 2x_2 - 3x_3 = -6 \\ x_1 + 4x_2 + 4x_3 = 3 \end{cases}$$

$$\textcircled{2} \begin{cases} -x_1 + x_2 + 3x_3 = 12 \\ 2x_1 - x_2 + 2x_3 = -8 \\ 4x_1 + x_2 - 4x_3 = 15 \end{cases}$$

$$\textcircled{3} \begin{cases} -2u - 4v + 3w = -1 \\ 2v - w = 1 \\ u + v - 3w = -6 \end{cases}$$

$$\textcircled{4} \begin{cases} -2x - y + 4t = 2 \\ 2x + 3y + 3z + 2t = 14 \\ x + 2y + z + t = 7 \\ -x - z + t = -1 \end{cases}$$

$$\textcircled{5} \begin{pmatrix} 6 & 1 & 1 \\ 2 & 4 & 0 \\ 1 & 2 & 6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 12 \\ 0 \\ 6 \end{pmatrix}$$

$$\textcircled{6} \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \\ 3 & 4 & 1 & 2 \\ 4 & 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 10 \\ 10 \\ 10 \\ 10 \end{pmatrix}$$

**Correction**

On utilise la méthode du pivot de GAUSS :

\textcircled{1}

$$\begin{cases} x_1 + 2x_2 - x_3 = 2, & L_2 \leftarrow L_2 - L_1 \\ x_1 - 2x_2 - 3x_3 = -6, & L_3 \leftarrow L_3 - L_1 \\ x_1 + 4x_2 + 4x_3 = 3. & \end{cases} \begin{cases} x_1 + 2x_2 - x_3 = 2, & L_3 \leftarrow L_3 + L_2 / 2 \\ -4x_2 - 2x_3 = -8, & \\ 2x_2 + 5x_3 = 1. & \end{cases} \begin{cases} x_1 + 2x_2 - x_3 = 2, & \\ -4x_2 - 2x_3 = -8, & \\ 4x_3 = -3. & \end{cases}$$

donc  $x_3 = \frac{-3}{4}$ ,  $x_2 = \frac{19}{8}$  et  $x_1 = \frac{-7}{2}$ .

\textcircled{2}

$$\begin{cases} -x_1 + x_2 + 3x_3 = 12 & L_2 \leftarrow L_2 + 2L_1 \\ 2x_1 - x_2 + 2x_3 = -8 & L_3 \leftarrow L_3 + 4L_1 \\ 4x_1 + x_2 - 4x_3 = 15 & \end{cases} \begin{cases} -x_1 + x_2 + 3x_3 = 12 & \\ x_2 + 8x_3 = 16 & \\ 5x_2 + 8x_3 = 63 & \end{cases} \begin{cases} -x_1 + x_2 + 3x_3 = 12 & \\ x_2 + 8x_3 = 16 & \\ -32x_3 = -17 & \end{cases}$$

donc  $x_3 = \frac{17}{32}$ ,  $x_2 = \frac{47}{4}$  et  $x_1 = \frac{43}{32}$ .

\textcircled{3}

$$\begin{cases} -2u - 4v + 3w = -1 & L_2 \leftarrow L_2 + L_1 / 2 \\ 2v - w = 1 & \\ u + v - 3w = -6 & \end{cases} \begin{cases} -2u - 4v + 3w = -1 & \\ 2v - w = 1 & \\ -v - \frac{3}{2}w = -13 / 2 & \end{cases} \begin{cases} -2u - 4v + 3w = -1 & \\ 2v - w = 1 & \\ -2w = -6 & \end{cases}$$

donc  $w = 3$ ,  $v = 2$  et  $u = 1$ .

\textcircled{4}

$$\begin{cases} -2x - y + 4t = 2 & L_2 \leftarrow L_2 + L_1 / 2 \\ 2x + 3y + 3z + 2t = 14 & L_3 \leftarrow L_3 + L_1 / 2 \\ x + 2y + z + t = 7 & L_4 \leftarrow L_4 - L_1 / 2 \\ -x - z + t = -1 & \end{cases} \begin{cases} -2x - y + 4t = 2 & \\ 2y + 3z + 6t = 16 & \\ \frac{3}{2}y + z + 3t = 8 & \\ \frac{1}{2}y - z - t = -2 & \end{cases}$$

$$\xrightarrow[L_3 \leftarrow L_3 - 3L_2/4]{L_4 \leftarrow L_2 - L_2/4} \left\{ \begin{array}{l} -2x-y+4t=2 \\ 2y+3z+6t=16 \\ -\frac{5}{4}z-\frac{3}{2}t=-4 \\ -\frac{7}{4}z-\frac{5}{2}t=-6 \end{array} \right. \xrightarrow[L_4 \leftarrow L_4 - 7L_3/5]{} \left\{ \begin{array}{l} -2x-y+4t=2 \\ 2y+3z+6t=16 \\ -\frac{5}{4}z-\frac{3}{2}t=-4 \\ -\frac{5}{5}z-\frac{3}{2}t=-\frac{2}{5} \end{array} \right.$$

donc  $t = 1$ ,  $z = 2$ ,  $y = 2$  et  $x = 0$ .

(5)

$$[\mathbb{A}|\mathbf{b}] = \left( \begin{array}{ccc|c} 6 & 1 & 1 & 12 \\ 2 & 4 & 0 & 0 \\ 1 & 2 & 6 & 6 \end{array} \right) \xrightarrow[L_2 \leftarrow L_2 - \frac{2}{6}L_1]{L_3 \leftarrow L_3 - \frac{1}{6}L_1} \left( \begin{array}{ccc|c} 6 & 1 & 1 & 12 \\ 0 & \frac{11}{6} & -\frac{1}{3} & -4 \\ 0 & \frac{11}{6} & \frac{35}{6} & 4 \end{array} \right) \xrightarrow[L_3 \leftarrow L_3 - \frac{\frac{11}{6}}{3}L_2]{} \left( \begin{array}{ccc|c} 6 & 1 & 1 & 12 \\ 0 & \frac{11}{3} & -\frac{1}{3} & -4 \\ 0 & 0 & 6 & 6 \end{array} \right)$$

donc

$$\begin{cases} 6x_1 + x_2 + x_3 = 12, \\ \frac{11}{3}x_2 - \frac{1}{3}x_3 = -4 \\ 6x_3 = 6 \end{cases} \implies x_3 = 1, \quad x_2 = -1, \quad x_1 = 2.$$

(6)

$$\begin{aligned} [\mathbb{A}|\mathbf{b}] = & \left( \begin{array}{cccc|c} 1 & 2 & 3 & 4 & 10 \\ 2 & 3 & 4 & 1 & 10 \\ 3 & 4 & 1 & 2 & 10 \\ 4 & 1 & 2 & 3 & 10 \end{array} \right) \xrightarrow[L_2 \leftarrow L_2 - 2L_1]{L_3 \leftarrow L_3 - 3L_1} \left( \begin{array}{cccc|c} 1 & 2 & 3 & 4 & 10 \\ 0 & -1 & -2 & -7 & -10 \\ 0 & -2 & -8 & -10 & -20 \\ 0 & -7 & -10 & -13 & -30 \end{array} \right) \\ & \xrightarrow[L_3 \leftarrow L_3 - 2L_2]{L_4 \leftarrow L_4 - 7L_2} \left( \begin{array}{cccc|c} 1 & 2 & 3 & 4 & 10 \\ 0 & -1 & -2 & -7 & -10 \\ 0 & 0 & -4 & 4 & 0 \\ 0 & 0 & 4 & 36 & 40 \end{array} \right) \xrightarrow[L_4 \leftarrow L_4 + L_3]{} \left( \begin{array}{cccc|c} 1 & 2 & 3 & 4 & 10 \\ 0 & -1 & -2 & -7 & -10 \\ 0 & 0 & -4 & 4 & 0 \\ 0 & 0 & 0 & 40 & 40 \end{array} \right) \end{aligned}$$

donc

$$\begin{cases} x_1 + 2x_2 + 3x_3 + 4x_4 = 10 \\ -x_2 - 2x_3 - 7x_4 = -10 \\ -4x_3 + 4x_4 = 0 \\ 40x_4 = 40 \end{cases} \implies x_4 = 1, \quad x_3 = 1, \quad x_2 = 1, \quad x_1 = 1.$$

```
A=[1 2 -1; 1 -2 -3; 1 4 4]
b=[2; -6; 3]
A\b
A=[-1 1 3; 2 -1 2; 4 1 -4]
b=[12; -8; 15]
A\b
A=[-2 -4 3; 0 2 -1; 1 1 -3]
b=[-1; 1; -6]
A\b
```

```
A=[-2 -1 0 4; 2 3 3 2; 1 2 1 1; -1 0 -1 1]
b=[2; 14; 7; -1]
A\b
A=[6 1 1; 2 4 0; 1 2 6]
b=[12; 0; 6]
A\b
A=[1 2 3 4; 2 3 4 1; 3 4 1 2; 4 1 2 3]
b=[10; 10; 10; 10]
A\b
```

## ★ Exercice 5.4

Considérons un système linéaire sous la forme matricielle  $\mathbb{A}\mathbf{x} = \mathbf{b}$  où  $\mathbb{A}$  est une matrice de  $\mathbb{R}^{n \times n}$  non singulière et  $\mathbf{b}$  est un vecteur colonne de  $\mathbb{R}^n$ .

Implémenter une fonction appelée `mygauss` qui transforme la matrice augmentée  $[\mathbb{A}|\mathbf{b}]$  en une matrice triangulaire supérieure par la méthode de GAUSS et, à chaque étape, affiche les opérations sur les lignes ainsi que la matrice modifiée. Enfin, elle résout le système linéaire triangulaire par remontée.

La syntaxe doit être `function [x]=mygauss(A,b)`

Écrire un script appelé `TESTmygauss.m` pour tester cette fonction sur l'exemple suivant : pour

$$\mathbb{A} = \begin{pmatrix} 1 & 0 & 3 \\ 2 & 2 & 2 \\ 3 & 6 & 4 \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} 4 \\ 6 \\ 13 \end{pmatrix}$$

on doit obtenir

$$\mathbf{x} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

### Correction

Dans le fichier mygauss.m on écrit

```
function [x]=mygauss(A,b)
printf("Matrice augmentee : [A|b]\n")
Ab = [A,b]
[n,m]=size(A);
tol=1.0e-9;
for k=1:n-1
    printf(strcat("\nEtape ",num2str(k),"\n"))
    for i=k+1:n
        L(i,k)=Ab(i,k)/Ab(k,k);
        printf(strcat("\tL_ ",num2str(i)," <- L_ ",num2str(i)," - (",num2str(L(i,k)),") L_ ",num2str(k),"\n"))
        Ab(i,k:n+1)=Ab(i,k:n+1)-L(i,k)*Ab(k,k:n+1);
    end
    Ab
end
printf("\nResolution du systeme triangulaire ainsi obtenu\n")
U=triu(Ab(:,1:n));
y=Ab(:,n+1);
x(n)=y(n)/U(n,n);
for i=n-1:-1:1
    x(i)=(y(i)-dot(U(i,i+1:n),x(i+1:n)))/U(i,i);
end
end
```

et on teste cette fonction par exemple comme suit

```
clear all
A=[1 0 3; 2 2 2; 3 6 4];
b=[4; 6; 13];
x=mygauss(A,b),
% Pour verifier notre resultat on peut
% comparer au resultat d'Octave
xOctave=A\b
% ou verifier que Ax=b
printf(strcat("||Ax-b||=",num2str(norm(A*x-b)), "\n"))
```

### Exercice 5.5

Soit le système linéaire

$$(S) \quad \begin{cases} 2x_1 - x_2 - 3x_3 = 0, \\ -x_1 + 2x_3 = 0, \\ 2x_1 - 3x_2 - x_3 = 0. \end{cases}$$

Ce système est-il compatible ? Possède-t-il une solution unique ?

### Correction

$$\begin{cases} 2x_1 - x_2 - 3x_3 = 0, \\ -x_1 + 2x_3 = 0, \\ 2x_1 - 3x_2 - x_3 = 0. \end{cases} \xrightarrow{\begin{array}{l} L_2 \leftarrow L_2 + L_1/2 \\ L_3 \leftarrow L_3 - L_1 \end{array}} \begin{cases} 2x_1 - x_2 - 3x_3 = 0, \\ -1/2x_2 + 1/2x_3 = 0, \\ -2x_2 + 2x_3 = 0, \end{cases} \xrightarrow{L_3 \leftarrow L_3 - 4L_2} \begin{cases} 2x_1 - x_2 - 3x_3 = 0, \\ -1/2x_2 + 1/2x_3 = 0, \\ 0 = 0, \end{cases}$$

Le système est compatible car le rang du système est 2 inférieur au nombre d'inconnues 3 et la solution n'est pas unique car  $\text{rg}(S) < 3$ . Il admet une infinité de solutions de la forme  $(2\kappa, \kappa, \kappa)$ ,  $\kappa \in \mathbb{R}$ .

**Exercice 5.6**

Trouver toutes les solutions du système linéaire homogène

$$(S) \quad \begin{cases} -3x_1 + x_2 + 2x_3 = 0, \\ -2x_1 + 2x_3 = 0, \\ -11x_1 + 6x_2 + 5x_3 = 0. \end{cases}$$

**Correction**

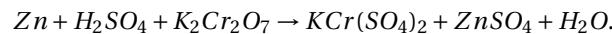
Le système étant homogène, il est inutile d'écrire le terme source dans la méthode du pivot de GAUSS :

$$\mathbb{A} = \begin{pmatrix} -3 & 1 & 2 \\ -2 & 0 & 2 \\ -11 & 6 & 5 \end{pmatrix} \xrightarrow{\substack{L_2 \leftarrow L_2 - 2L_1/3 \\ L_3 \leftarrow L_3 - 11L_1/3}} \begin{pmatrix} -3 & 1 & 2 \\ 0 & -2/3 & 2/3 \\ 0 & 7/3 & -7/3 \end{pmatrix} \xrightarrow{L_3 \leftarrow L_3 + 7L_2/2} \begin{pmatrix} -3 & 1 & 2 \\ 0 & -2/3 & 2/3 \\ 0 & 0 & 0 \end{pmatrix}$$

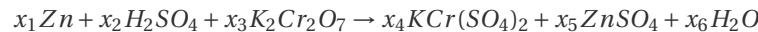
Le système admet une infinité de solutions de la forme  $(\kappa, \kappa, \kappa)$  avec  $\kappa \in \mathbb{R}$ .

**Exercice 5.7**

Équilibrer la réaction

**Correction**

Écrivons les coefficients stœchiométriques et les contraintes :



1. Atomes de  $Zn$  :  $x_1 = x_5$ , i.e.  $x_1 - x_5 = 0$
2. Atomes de  $H$  :  $2x_2 = 2x_6$ , i.e.  $x_2 - x_6 = 0$
3. Atomes de  $S$  :  $x_2 = 2x_4 + x_5$ , i.e.  $x_2 - 2x_4 - x_5 = 0$
4. Atomes de  $K$  :  $2x_3 = x_4$ , i.e.  $2x_3 - x_4 = 0$
5. Atomes de  $Cr$  :  $2x_3 = x_4$ , i.e.  $2x_3 - x_4 = 0$
6. Atomes de  $O$  :  $4x_2 + 7x_3 = 8x_4 + 4x_5 + x_6$ , i.e.  $4x_2 + 7x_3 - 8x_4 - 4x_5 - x_6 = 0$

Notons que la contrainte  $2x_3 - x_4 = 0$  est répétée deux fois, donc on ne l'écrira qu'une seule fois dans le système linéaire ; cela donne 5 équations pour 6 inconnues. Fixons arbitrairement un des coefficients, par exemple  $x_6 = 1$  ; on obtient alors le système linéaire

$$\begin{pmatrix} 1 & 0 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -2 & -1 \\ 0 & 0 & 2 & -1 & 0 \\ 0 & 4 & 7 & -8 & -4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

ce qui donne

$$\begin{array}{c} \left( \begin{array}{ccccc|c} 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & -2 & -1 & 0 \\ 0 & 0 & 2 & -1 & 0 & 0 \\ 0 & 4 & 7 & -8 & -4 & 1 \end{array} \right) \xrightarrow{\substack{L_3 \leftarrow L_3 - L_2 \\ L_5 \leftarrow L_5 - 4L_2}} \left( \begin{array}{ccccc|c} 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -2 & -1 & -1 \\ 0 & 0 & 2 & -1 & 0 & 0 \\ 0 & 0 & 7 & -8 & -4 & -3 \end{array} \right) \xrightarrow{L_3 \leftrightarrow L_4} \left( \begin{array}{ccccc|c} 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 2 & -1 & 0 & 0 \\ 0 & 0 & 0 & -2 & -1 & -1 \\ 0 & 0 & 7 & -8 & -4 & -3 \end{array} \right) \\ \xrightarrow{L_5 \leftarrow L_5 - \frac{7}{2}L_3} \left( \begin{array}{ccccc|c} 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 2 & -1 & 0 & 0 \\ 0 & 0 & 0 & -2 & -1 & -1 \\ 0 & 0 & 0 & -\frac{9}{2} & -4 & -3 \end{array} \right) \xrightarrow{L_5 \leftarrow L_5 - \frac{9}{4}L_4} \left( \begin{array}{ccccc|c} 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 2 & -1 & 0 & 0 \\ 0 & 0 & 0 & -2 & -1 & -1 \\ 0 & 0 & 0 & 0 & -\frac{7}{4} & -\frac{3}{4} \end{array} \right) \end{array}$$

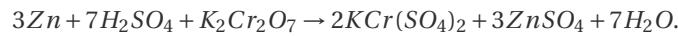
dont la solution est bien

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 3/7 \\ 1 \\ 1/7 \\ 2/7 \\ 3/7 \end{pmatrix}$$

Si on multiplie tous les coefficients par 7 on obtient

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 3 \\ 7 \\ 1 \\ 2 \\ 3 \\ 7 \end{pmatrix}$$

et donc la réaction équilibrée



### Exercice 5.8 (V. GUIRARDEL)

Vous projetez de passer un concours de recrutement l'an prochain. Vous avez sous les yeux le tableau de notes suivant :

CANDIDAT	Mathématique	Anglais	Informatique	Moyenne
QUI	7	12	6	8
QUO	11	6	10	9
QUA	11	16	14	14

Retrouver les coefficients de chaque épreuve. La solution est-elle unique ?

#### Correction

Il s'agit de trouver les trois coefficients  $m, a, i \in [0; 1]$  tels que

$$\begin{cases} 7m + 12a + 6i = 8, \\ 11m + 6a + 10i = 9, \\ 11m + 16a + 14i = 14. \end{cases}$$

Utilisons la méthode de GAUSS :

$$\left\{ \begin{array}{l} 7m+12a+6i=8, \\ 11m+6a+10i=9, \\ 11m+16a+14i=14, \end{array} \right. \xrightarrow{\begin{array}{l} L_2 \leftarrow L_2 - \frac{11}{7}L_1 \\ L_3 \leftarrow L_3 - \frac{11}{7}L_1 \end{array}} \left\{ \begin{array}{l} 7m+12a+6i=8, \\ -\frac{90}{7}a+\frac{4}{7}i=-\frac{25}{7}, \\ -\frac{20}{7}a+\frac{32}{7}i=\frac{10}{7}, \end{array} \right. \xrightarrow{\begin{array}{l} L_3 \leftarrow L_3 - \frac{2}{9}L_2 \\ -\frac{90}{7}a+\frac{4}{7}i=-\frac{25}{7}, \\ \frac{40}{9}i=\frac{20}{9}, \end{array}} \left\{ \begin{array}{l} 7m+12a+6i=8, \\ -\frac{90}{7}a+\frac{4}{7}i=-\frac{25}{7}, \\ \frac{40}{9}i=\frac{20}{9}, \end{array} \right. \right.$$

qui admet l'unique solution  $(0.2, 0.3, 0.5)$ .

Une autre interprétation est la suivante : il s'agit de trouver les trois coefficients  $m, a, i \in [0; 1]$  tels que

$$\begin{cases} 7m + 12a + 6i = 8(m + a + i) \\ 11m + 6a + 10i = 9(m + a + i), \\ 11m + 16a + 14i = 14(m + a + i). \end{cases}$$

Utilisons la méthode de GAUSS :

$$\left\{ \begin{array}{l} -m+4a-2i=0, \\ 2m-3a+i=0, \\ -3m+2a=0, \end{array} \right. \xrightarrow{\begin{array}{l} L_2 \leftarrow L_2 - \frac{11}{7}L_1 \\ L_3 \leftarrow L_3 - \frac{11}{7}L_1 \end{array}} \left\{ \begin{array}{l} -m+4a-2i=0, \\ 5a-3i=0, \\ -10a+6i=0, \end{array} \right. \xrightarrow{\begin{array}{l} L_3 \leftarrow L_3 - \frac{2}{9}L_2 \\ 5a-3i=0, \\ 0=0, \end{array}} \left\{ \begin{array}{l} -m+4a-2i=0, \\ 5a-3i=0, \\ 0=0, \end{array} \right. \right.$$

qui admet une infinité de solutions de la forme  $(2\kappa, 3\kappa, 5\kappa)$  avec  $\kappa \in [0; 1/5]$ .

### Exercice 5.9 (V. GUIRARDEL)

Une entreprise fabrique des manteaux. Ces manteaux sont composés de tissu rouge, de tissu bleu et d'une doublure noire. Le tableau suivant résume les mètres carrés de chaque tissu nécessaires à la confection du manteau en tailles S, M, L et XL :

	S	M	L	XL
Tissu rouge	0.4	0.5	0.6	0.7
Tissu bleu	1	1.1	1.2	1.3
Doublure	1.5	1.7	1.9	2.1

Chaque tissu est tissé à l'aide de plusieurs types de fil : coton, polyester et polyamide. Le tableau suivant résume les mètres de fil de chaque type nécessaires par mètre carré de tissu :

	Tissu rouge	Tissu bleu	Doublure
Coton	500	400	1000
Polyamide	1000	900	700
Polyester	500	600	0

1. L'entreprise veut produire  $s$  manteaux taille S,  $m$  manteaux taille M,  $\ell$  manteaux taille L et  $x$  manteaux taille XL. Quelle quantité de fil de chaque catégorie doit-elle commander ? Répondre à cette question dans le langage des matrices.
2. En fin d'année, l'entreprise veut écouler entièrement ses stocks de fils. Il lui reste 100 000 m de coton et de polyamide, et 20 000 m de Polyester. Peut-elle transformer entièrement ses stocks de fils en manteaux ?

### Correction

Introduisons les deux matrices  $\mathbb{A}$  et  $\mathbb{B}$  et les deux vecteurs  $\mathbf{u}$  et  $\mathbf{v}$  suivants

$$\mathbb{A} = \begin{pmatrix} 0.4 & 0.5 & 0.6 & 0.7 \\ 1 & 1.1 & 1.2 & 1.3 \\ 1.5 & 1.7 & 1.9 & 2.1 \end{pmatrix} \quad \mathbb{B} = \begin{pmatrix} 500 & 400 & 1000 \\ 1000 & 900 & 700 \\ 500 & 600 & 0 \end{pmatrix} \quad \mathbf{u} = \begin{pmatrix} s \\ m \\ \ell \\ x \end{pmatrix} \quad \mathbf{v} = \begin{pmatrix} c \\ a \\ e \\ x \end{pmatrix}$$

1. Pour produire  $s$  manteaux taille S,  $m$  manteaux taille M,  $\ell$  manteaux taille L et  $x$  manteaux taille XL, l'entreprise doit commander  $c$  mètres de coton,  $a$  mètres de polyamide et  $e$  mètres de polyester où  $c, a, e$  sont les entrées du vecteur  $\mathbf{v}$  suivant :

$$\mathbf{v} = \mathbb{B}\mathbb{A}\mathbf{u} = \begin{pmatrix} 2100s + 2390m + 2680\ell + 2970x \\ 2350s + 2680m + 3010\ell + 3340x \\ 800s + 910m + 1020\ell + 1130x \end{pmatrix}.$$

2. On cherche s'il existe un vecteur  $\mathbf{u}$  tel que

$$\begin{pmatrix} 100000 \\ 100000 \\ 20000 \end{pmatrix} = \mathbb{B}\mathbb{A}\mathbf{u},$$

i.e. s'il existe une solution du système linéaire

$$\begin{pmatrix} 2100 & 2390 & 2680 & 2970 \\ 2350 & 2680 & 3010 & 3340 \\ 800 & 910 & 1020 & 1130 \end{pmatrix} \begin{pmatrix} s \\ m \\ \ell \\ x \end{pmatrix} = \begin{pmatrix} 100000 \\ 100000 \\ 20000 \end{pmatrix}.$$

En appliquant la méthode de GAUSS on obtient le système

$$\begin{pmatrix} 2100 & 2390 & 2680 & 2970 \\ 0 & \frac{115}{21} & \frac{230}{21} & \frac{115}{7} \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} s \\ m \\ \ell \\ x \end{pmatrix} = \begin{pmatrix} 100000 \\ -\frac{250000}{21} \\ -\frac{440000}{23} \end{pmatrix}$$

qui n'admet pas de solution.

### Exercice 5.10

Soit le système linéaire

$$(S) \quad \begin{cases} x - \alpha y = 1, \\ \alpha x - y = 1. \end{cases}$$

Déterminer les valeurs de  $\alpha$  de telle sorte que ce système possède :

1. une infinité de solutions ;
2. aucune solution ;
3. une solution unique.

### Correction

$$\left( \begin{array}{cc|c} 1 & -\alpha & 1 \\ \alpha & -1 & 1 \end{array} \right) \xrightarrow{L_2 \leftarrow L_2 - \alpha L_1} \left( \begin{array}{cc|c} 1 & -\alpha & 1 \\ 0 & -1 + \alpha^2 & 1 - \alpha \end{array} \right).$$

Comme  $-1 + \alpha^2 = (\alpha - 1)(\alpha + 1)$  on conclut que

1. si  $\alpha = 1$  (i.e. la dernière équation correspond à  $0 = 0$ ) alors ( $S$ ) possède une infinité de solutions,
2. si  $\alpha = -1$  (i.e. la dernière équation correspond à  $0 = 2$ ) alors ( $S$ ) ne possède aucune solution,
3. si  $\alpha \notin \{-1; 1\}$  alors ( $S$ ) possède une solution unique  $x = \frac{1}{\alpha+1}$  et  $y = -\frac{1}{\alpha+1}$ .

### Exercice 5.11

Soit le système linéaire

$$(S) \quad \begin{cases} x + \alpha y = 1, \\ -\alpha x - y = 1. \end{cases}$$

En utilisant le pivot de GAUSS, déterminer les valeurs de  $\alpha \in \mathbb{R}$  de telle sorte que ce système possède :

- a) une infinité de solutions ;
- b) aucune solution ;
- c) une solution unique.

#### Correction

$$\left\{ \begin{array}{l} x + \alpha y = 1 \\ -\alpha x - y = 1 \end{array} \right. \xrightarrow{L_2 \leftarrow L_2 + \alpha L_1} \left\{ \begin{array}{l} x + \alpha y = 1 \\ (-1 + \alpha^2)y = \alpha + 1 \end{array} \right.$$

Comme  $-1 + \alpha^2 = (\alpha - 1)(\alpha + 1)$  on conclut que

- a) si  $\alpha = -1$  (i.e. la dernière équation correspond à  $0 = 0$ ) alors ( $S$ ) possède une infinité de solutions,
- b) si  $\alpha = 1$  (i.e. la dernière équation correspond à  $0 = 2$ ) alors ( $S$ ) ne possède aucune solution,
- c) si  $\alpha \notin \{-1; 1\}$  alors ( $S$ ) possède une solution unique  $x = -\frac{1}{\alpha-1}$  et  $y = \frac{1}{\alpha-1}$ .

### Exercice 5.12

Soit le système linéaire

$$(S) \quad \begin{cases} x + y - z = 1, \\ 2x + 3y + \beta z = 3, \\ x + \beta y + 3z = -3. \end{cases}$$

Déterminer les valeurs de  $\beta$  de telle sorte que ce système possède :

1. une infinité de solutions ;
2. aucune solution ;
3. une solution unique.

#### Correction

$$\left( \begin{array}{ccc|c} 1 & 1 & -1 & 1 \\ 2 & 3 & \beta & 3 \\ 1 & \beta & 3 & -3 \end{array} \right) \xrightarrow{\substack{L_2 \leftarrow L_2 - 2L_1 \\ L_3 \leftarrow L_3 - L_1}} \left( \begin{array}{ccc|c} 1 & 1 & -1 & 1 \\ 0 & 1 & \beta + 2 & 1 \\ 0 & \beta - 1 & 4 & -4 \end{array} \right) \xrightarrow{L_3 \leftarrow L_3 + (1-\beta)L_2} \left( \begin{array}{ccc|c} 1 & 1 & -1 & 1 \\ 0 & 1 & \beta + 2 & 1 \\ 0 & 0 & (6 - \beta - \beta^2) & -(3 + \beta) \end{array} \right).$$

Comme  $6 - \beta - \beta^2 = (2 - \beta)(3 + \beta)$  on conclut que

1. si  $\beta = -3$  (i.e. la dernière équation correspond à  $0z = 0$ ) alors ( $S$ ) possède une infinité de solutions,
2. si  $\beta = 2$  (i.e. la dernière équation correspond à  $0z = -5$ ) alors ( $S$ ) ne possède aucune solution,
3. si  $\beta \notin \{2; -3\}$  alors ( $S$ ) possède une solution unique.

### Exercice 5.13

Trouver les valeurs de  $\kappa \in \mathbb{R}$  pour lesquelles le système suivant a un nombre respectivement fini et infini de solutions :

$$\begin{cases} 2x_1 - x_2 = \kappa, \\ x_1 - x_2 - x_3 = 0, \\ x_1 - \kappa x_2 + \kappa x_3 = \kappa. \end{cases}$$

**Correction**

$$\left( \begin{array}{ccc|c} 2 & -1 & 0 & \kappa \\ 1 & -1 & -1 & 0 \\ 1 & -\kappa & \kappa & \kappa \end{array} \right) \xrightarrow{L_2 \leftarrow L_2 - L_1/2} \left( \begin{array}{ccc|c} 2 & -1 & 0 & \kappa \\ 0 & -1/2 & -1 & -\kappa/2 \\ 0 & -\kappa + 1/2 & \kappa & \kappa/2 \end{array} \right) \xrightarrow{L_3 \leftarrow L_3 + (1-2\kappa)L_2} \left( \begin{array}{ccc|c} 2 & -1 & 0 & \kappa \\ 0 & -1/2 & -1 & -\kappa/2 \\ 0 & 0 & 3\kappa - 1 & \kappa^2 \end{array} \right).$$

On conclut que

1. si  $\kappa = \frac{1}{3}$  alors ( $S$ ) ne possède aucune solution,
2. si  $\kappa \neq \frac{1}{3}$  alors ( $S$ ) possède une solution unique donnée par  $x_3 = \frac{\kappa^2}{3\kappa-1}$ ,  $x_2 = \frac{-\kappa/2+x_3}{-1/2} = \frac{\kappa(\kappa-1)}{3\kappa-1}$  et  $x_1 = \frac{\kappa+x_2}{2} = \frac{\kappa(2\kappa-1)}{3\kappa-1}$ ,
3. il n'existe aucune valeur de  $\kappa$  pour que ( $S$ ) possède une infinité de solutions.

**Exercice 5.14**

Résoudre le système linéaire en discutant suivant la valeur du paramètre  $a \in \mathbb{R}$ :

$$\begin{cases} x + 2y + 3z = 2, \\ x - y + 2z = 7, \\ 3x + az = 10. \end{cases}$$

**Correction**

Si on utilise la méthode de GAUSS on trouve

$$[\mathbb{A}|\mathbf{b}] = \left( \begin{array}{ccc|c} 1 & 2 & 3 & 2 \\ 1 & -1 & 2 & 7 \\ 3 & 0 & a & 10 \end{array} \right) \xrightarrow{L_2 \leftarrow L_2 - L_1} \left( \begin{array}{ccc|c} 1 & 2 & 3 & 2 \\ 0 & -3 & -1 & 5 \\ 3 & 0 & a-9 & 4 \end{array} \right) \xrightarrow{L_3 \leftarrow L_3 - 3L_1} \left( \begin{array}{ccc|c} 1 & 2 & 3 & 2 \\ 0 & -3 & -1 & 5 \\ 0 & 0 & a-7 & -6 \end{array} \right).$$

On a ainsi transformé le système linéaire initial dans le système linéaire triangulaire supérieur équivalent

$$\begin{cases} x + 2y + 3z = 2, \\ -3y - z = 5, \\ (a-7)z = -6. \end{cases}$$

Par conséquent,

- \* si  $a \neq 7$ ,  $z = \frac{-6}{a-7}$ ,  $y = \frac{5+z}{-3} = \frac{5a-41}{-3(a-7)}$  et  $x = 2 - 2y - 3z = \frac{2(8a-35)}{3(a-7)}$  est l'unique solution du système linéaire ;
- \* si  $a = 7$  il n'y a pas de solutions du système linéaire.

Observons que si on ne veut pas calculer la solution mais juste dire s'il en existe une (ou plusieurs), il suffit de regarder le rang des matrices  $\mathbb{A}$  et  $[\mathbb{A}|\mathbf{b}]$  :

- \*  $\text{rg}(\mathbb{A}) = \begin{cases} 3 & \text{si } a \neq 7 \\ 2 & \text{si } a = 7 \end{cases}$  car  $\det(\mathbb{A}) = 21 - 3a$  et  $\det(\mathbb{A}_{33}) \neq 0$  où  $\mathbb{A}_{33}$  est la sous-matrice de  $\mathbb{A}$  obtenue en supprimant la 3-ème ligne et la 3-ème colonne ;
- \*  $\text{rg}([\mathbb{A}|\mathbf{b}]) = 3$  car  $\det\left(\begin{smallmatrix} 1 & 2 & 2 \\ 1 & -1 & 7 \\ 3 & 0 & 10 \end{smallmatrix}\right) \neq 0$  où  $\left(\begin{smallmatrix} 1 & 2 & 2 \\ 1 & -1 & 7 \\ 3 & 0 & 10 \end{smallmatrix}\right)$  est la sous-matrice de  $[\mathbb{A}|\mathbf{b}]$  obtenue en supprimant la 3-ème colonne.

**Exercice 5.15**

En utilisant la méthode de GAUSS, résoudre le système linéaire en discutant suivant la valeur du paramètre  $a \in \mathbb{R}$ :

$$\begin{cases} x - y + 2z = 7, \\ x + 2y + 3z = 2, \\ 3x + az = 10. \end{cases}$$

**Correction**

$$\left( \begin{array}{ccc|c} 1 & -1 & 2 & 7 \\ 1 & 2 & 3 & 2 \\ 3 & 0 & a & 10 \end{array} \right) \xrightarrow{L_2 \leftarrow L_2 - L_1} \left( \begin{array}{ccc|c} 1 & -1 & 2 & 7 \\ 0 & 3 & 1 & -5 \\ 3 & 0 & a-6 & -11 \end{array} \right) \xrightarrow{L_3 \leftarrow L_3 - 3L_1} \left( \begin{array}{ccc|c} 1 & -1 & 2 & 7 \\ 0 & 3 & 1 & -5 \\ 0 & 0 & a-7 & -6 \end{array} \right).$$

On a ainsi transformé le système linéaire initial dans le système linéaire triangulaire supérieur équivalent

$$\begin{cases} x - y + 2z = 7, \\ 3y + z = -5, \\ (a - 7)z = -6. \end{cases}$$

Par conséquent,

- \* si  $a \neq 7$ ,  $z = \frac{-6}{a-7}$ ,  $y = \frac{-5-z}{3} = \frac{-5a+41}{3(a-7)}$  et  $x = 7 - y - 2z = \frac{2(8a-35)}{3(a-7)}$  est l'unique solution du système linéaire;
- \* si  $a = 7$  il n'y a pas de solutions du système linéaire.

Observons que si on ne veut pas calculer la solution mais juste dire s'il en existe une (ou plusieurs), il suffit de regarder le rang des matrices  $\mathbb{A}$  et  $[\mathbb{A}|\mathbf{b}]$  :

- \*  $\text{rg}(\mathbb{A}) = \begin{cases} 3 & \text{si } a \neq 7 \\ 2 & \text{si } a = 7 \end{cases}$  car  $\det(\mathbb{A}) = 3a - 21$  et  $\det(\mathbb{A}_{33}) \neq 0$  où  $\mathbb{A}_{33}$  est la sous-matrice de  $\mathbb{A}$  obtenue en supprimant la 3-ème ligne et la 3-ème colonne;
- \*  $\text{rg}([\mathbb{A}|\mathbf{b}]) = 3$  car  $\det\left(\begin{pmatrix} 1 & 2 & 2 \\ 1 & -1 & 7 \\ 3 & 0 & 10 \end{pmatrix}\right) \neq 0$  où  $\begin{pmatrix} 1 & 2 & 2 \\ 1 & -1 & 7 \\ 3 & 0 & 10 \end{pmatrix}$  est la sous-matrice de  $[\mathbb{A}|\mathbf{b}]$  obtenue en supprimant la 3-ème colonne.

### Exercice 5.16

En utilisant la méthode du pivot de GAUSS, résoudre le système linéaire en discutant suivant la valeur du paramètre  $a \in \mathbb{R}$ :

$$\begin{cases} x + z + w = 0, \\ ax + y + (a-1)z + w = 0, \\ 2x + ay + z + 2w = 0, \\ x - y + 2z + aw = 0. \end{cases}$$

#### Correction

Il s'agit d'un système homogène, il est alors inutile d'écrire le terme source dans la méthode du pivot de GAUSS. En appliquant cette méthode on obtient

$$\left( \begin{array}{cccc} 1 & 0 & 1 & 1 \\ a & 1 & a-1 & 1 \\ 2 & a & 1 & 2 \\ 1 & -1 & 2 & a \end{array} \right) \xrightarrow[L_3 \leftarrow L_3 - 2L_1]{L_2 \leftarrow L_2 - aL_1} \left( \begin{array}{cccc} 1 & 0 & 1 & 1 \\ 0 & 1 & -1 & 1-a \\ 0 & a & -1 & 0 \\ 0 & -1 & 1 & a-1 \end{array} \right) \xrightarrow[L_4 \leftarrow L_4 + L_2]{L_3 \leftarrow L_3 - aL_2} \left( \begin{array}{cccc} 1 & 0 & 1 & 1 \\ 0 & 1 & -1 & 1-a \\ 0 & 0 & a-1 & a(a-1) \\ 0 & 0 & 0 & 0 \end{array} \right).$$

On a ainsi transformé le système linéaire initial dans le système linéaire triangulaire supérieur équivalent

$$\begin{cases} x + z + w = 0, \\ y - z + (1-a)w = 0, \\ (a-1)z + a(a-1)w = 0, \\ 0 = 0. \end{cases}$$

Par conséquent, si on pose  $w = \kappa_1 \in \mathbb{R}$  une constante réelle quelconque, alors

- \* si  $a \neq 1$ ,  $z = \frac{-a(a-1)w}{a-1} = -a\kappa_1$ ,  $y = -(1-a)w + z = -\kappa_1$  et  $x = -w - z = (a-1)\kappa_1$ : tous les vecteurs de  $\text{Vect}\{(a-1, -1, -a, 1)\}$  sont solution du système linéaire;
- \* si  $a = 1$ , on pose  $z = \kappa_2 \in \mathbb{R}$  une constante réelle quelconque et on a  $y = -(1-a)w + z = -(1-a)\kappa_1 + \kappa_2$  et  $x = -w - z = -\kappa_1 - \kappa_2$ : tous les vecteurs de  $\text{Vect}\{(-1, 1, 1, 0), (-1, 0, 0, 1)\}$  sont solution du système linéaire.

### Exercice 5.17

En utilisant la méthode du pivot de GAUSS, résoudre le système linéaire en discutant suivant la valeur du paramètre  $b \in \mathbb{R}$ :

$$\begin{cases} x + z + w = 0, \\ (b+1)x + y + bz + w = 0, \\ 2x + (b+1)y + z + 2w = 0, \\ x - y + 2z + (b+1)w = 0. \end{cases}$$

**Correction**

Il s'agit d'un système homogène, il est alors inutile d'écrire le terme source dans la méthode du pivot de GAUSS. En appliquant cette méthode on obtient

$$\left( \begin{array}{cccc} 1 & 0 & 1 & 1 \\ b+1 & 1 & b & 1 \\ 2 & b+1 & 1 & 2 \\ 1 & -1 & 2 & b+1 \end{array} \right) \xrightarrow{\begin{array}{l} L_2 \leftarrow L_2 - (b+1)L_1 \\ L_3 \leftarrow L_3 - 2L_1 \\ L_4 \leftarrow L_4 - L_1 \end{array}} \left( \begin{array}{cccc} 1 & 0 & 1 & 1 \\ 0 & 1 & -1 & -b \\ 0 & b+1 & -1 & 0 \\ 0 & -1 & 1 & b \end{array} \right) \xrightarrow{\begin{array}{l} L_3 \leftarrow L_3 - (b+1)L_2 \\ L_4 \leftarrow L_4 + L_2 \end{array}} \left( \begin{array}{cccc} 1 & 0 & 1 & 1 \\ 0 & 1 & -1 & -b \\ 0 & 0 & b & b(b+1) \\ 0 & 0 & 0 & 0 \end{array} \right).$$

On a ainsi transformé le système linéaire initial dans le système linéaire triangulaire supérieur équivalent

$$\begin{cases} x + z + w = 0, \\ y - z - bw = 0, \\ bz + b(b+1)w = 0, \\ 0 = 0. \end{cases}$$

Par conséquent, si on pose  $w = \kappa_1 \in \mathbb{R}$  une constante réelle quelconque, alors

- \* si  $b \neq 0$ ,

$$z = \frac{-b(b+1)w}{b} = -(b+1)\kappa_1, \quad y = bw + z = -\kappa_1, \quad x = -w - z = b\kappa_1;$$

tous les vecteurs de  $\text{Vect}\{(b+1, 1, b+1, -1)\}$  sont solution du système linéaire ;

- \* si  $b = 0$ , on pose  $z = \kappa_2 \in \mathbb{R}$  une constante réelle quelconque et on a  $y = bw + z = \kappa_2$  et  $x = -w - z = -\kappa_1 - \kappa_2$  : tous les vecteurs de  $\text{Vect}\{(-1, 1, 1, 0), (-1, 0, 0, 1)\}$  sont solution du système linéaire.

**Exercice 5.18**

Discuter et résoudre le système

$$(S_m) \quad \begin{cases} (4m^2 - 1)x + (2m - 1)^2y = (2m + 1)^2, \\ (2m + 1)x + (4m - 1)y = 4m^2 - 1, \end{cases}$$

d'inconnues  $x, y \in \mathbb{R}$  et de paramètre  $m \in \mathbb{R}$ .

**Correction**

Puisque le système contient un paramètre, on commence par calculer le déterminant de la matrice  $\mathbb{A}$  :

$$\begin{vmatrix} 4m^2 - 1 & (2m - 1)^2 \\ 2m + 1 & 4m - 1 \end{vmatrix} = (4m^2 - 1)(4m - 1) - (2m - 1)^2(2m + 1) = 2m(2m - 1)(2m + 1).$$

On a

$$\text{rg}(\mathbb{A}) = \begin{cases} 2 & \text{si } m \in \mathbb{R} \setminus \{-\frac{1}{2}, 0, \frac{1}{2}\}, \\ 1 & \text{si } m \in \{-\frac{1}{2}, 0, \frac{1}{2}\}. \end{cases}$$

Pour calculer le rang de la matrice augmentée, on remarque que

$$\text{rg}([\mathbb{A}|\mathbf{b}]) = \begin{cases} 2 & \text{si } m \in \mathbb{R} \setminus \{-\frac{1}{2}, 0, \frac{1}{2}\}, \\ 1 \text{ ou } 2 & \text{si } m \in \{-\frac{1}{2}, 0, \frac{1}{2}\}. \end{cases}$$

Par conséquent :

- \* si  $m \in \mathbb{R} \setminus \{-\frac{1}{2}, 0, \frac{1}{2}\}$ , on a  $\text{rg}(\mathbb{A}) = \text{rg}([\mathbb{A}|\mathbf{b}]) = 2$  et on a deux inconnues donc le système est de CRAMER (*i.e.* il admet une et une seule solution),
- \* si  $m \in \{-\frac{1}{2}, 0, \frac{1}{2}\}$  il faut étudier chaque cas séparément sachant que
  - \* si  $\text{rg}(\mathbb{A}) = \text{rg}([\mathbb{A}|\mathbf{b}]) = 1$ , comme on a deux inconnues alors il y a une infinité de solution,
  - \* si  $\text{rg}(\mathbb{A}) = 1$  et  $\text{rg}([\mathbb{A}|\mathbf{b}]) = 2$ , alors le système n'as pas de solutions.

Étudions donc chaque cas :

- \* *Étude du cas*  $m = -\frac{1}{2}$ . Le système s'écrit

$$(S_{-\frac{1}{2}}) \quad \begin{cases} 4y = 0, \\ -3y = 0, \end{cases} \iff \begin{cases} x \in \mathbb{R}, \\ y = 0. \end{cases}$$

\* Étude du cas  $m = 0$ . Le système s'écrit

$$(S_0) \quad \begin{cases} -x + y = 1, \\ x - y = -1, \end{cases} \iff \begin{cases} y \in \mathbb{R}, \\ x = -1 + y. \end{cases}$$

\* Étude du cas  $m = \frac{1}{2}$ . Le système s'écrit

$$(S_{1/2}) \quad \begin{cases} 0 = 4, \\ 2x + y = 0, \end{cases}$$

qui n'admet pas de solutions.

\* Étude du cas  $m \in \mathbb{R} \setminus \{-\frac{1}{2}, 0, \frac{1}{2}\}$ . On peut utiliser la méthode de CRAMER : l'unique solution est donnée par

$$x = \frac{1}{2m(2m-1)(2m+1)} \begin{vmatrix} (2m+1)^2 & (2m-1)^2 \\ 4m^2-1 & 4m-1 \end{vmatrix} = \frac{-2(2m^2-5m+1)}{2m-1},$$

$$y = \frac{1}{2m(2m-1)(2m+1)} \begin{vmatrix} 4m^2-1 & (2m+1)^2 \\ 2m+1 & 4m^2-1 \end{vmatrix} = \frac{(2m+1)(2m-3)}{2m-1}.$$

Donc si on note  $\mathcal{S}$  l'ensemble des solutions,

$$\mathcal{S} = \begin{cases} \{(x, 0) \mid x \in \mathbb{R}\} & \text{si } m = -\frac{1}{2}, \\ \{(-1+y, y) \mid y \in \mathbb{R}\} & \text{si } m = 0, \\ \emptyset & \text{si } m = \frac{1}{2}, \\ \left\{ \left( \frac{-2(2m^2-5m+1)}{2m-1}, \frac{(2m+1)(2m-3)}{2m-1} \right) \right\} & \text{sinon.} \end{cases}$$

### Exercice 5.19

Discuter et résoudre le système

$$(S_a) \quad \begin{cases} (1+a)x + y + z = 0, \\ x + (1+a)y + z = 0, \\ x + y + (1+a)z = 0, \end{cases}$$

d'inconnue  $(x, y, z) \in \mathbb{R}^3$  et de paramètre  $a \in \mathbb{R}$ .

#### Correction

Comme le système contient un paramètre, on commence par calculer le déterminant de la matrice associée :

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

$$= ((1+a)-1)((1+a)^2 + (1+a) - 2) = ((1+a)-1)((1+a)+2)((1+a)-1) = a^2(3+a).$$

Le système est de Cramer si et seulement si ce déterminant est non nul, donc

$$(S_a) \text{ est de Cramer si et seulement } a \in \mathbb{R} \setminus \{-3, 0\}.$$

Notons  $\mathcal{S}$  l'ensemble des solutions.

\* Étude du cas  $a = -3$ . Le système s'écrit

$$(S_{-3}) \quad \begin{cases} -2x + y + z = 0, \\ x - 2y + z = 0, \\ x + y - 2z = 0, \end{cases}$$

On utilise la méthode du pivot de GAUSS :

$$\begin{cases} -2x + y + z = 0 \\ x - 2y + z = 0 \\ x + y - 2z = 0 \end{cases} \xrightarrow{\begin{array}{l} L_2 \leftarrow L_2 + L_1/2 \\ L_3 \leftarrow L_3 + L_1/2 \end{array}} \begin{cases} -2x + y + z = 0 \\ -\frac{3}{2}y + \frac{3}{2}z = 0 \\ \frac{3}{2}y - \frac{3}{2}z = 0 \end{cases} \xrightarrow{L_3 \leftarrow L_3 + L_2} \begin{cases} -2x - y + z = 0 \\ -\frac{5}{2}y + \frac{3}{2}z = 0 \\ 0z = 0 \end{cases}$$

donc  $z = \kappa \in \mathbb{R}$ ,  $y = z$  et  $x = z$ , ainsi

$$\mathcal{S} = \{(\kappa, \kappa, \kappa) \mid \kappa \in \mathbb{R}\}.$$

- ★ Étude du cas  $a = 0$ . Le système s'écrit

$$(S_0) \quad \begin{cases} x + y + z = 0, \\ x + y + z = 0, \\ x + y + z = 0, \end{cases}$$

donc  $z = \kappa_1 \in \mathbb{R}$ ,  $y = \kappa_2 \in \mathbb{R}$  et  $x = -\kappa_1 - \kappa_2$ , ainsi

$$\mathcal{S} = \{(-\kappa_1 - \kappa_2, \kappa_2, \kappa_1) \mid (\kappa_1, \kappa_2) \in \mathbb{R}^2\}.$$

- ★ Étude du cas  $a \in \mathbb{R} \setminus \{-3, 0\}$ . Il s'agit d'un système de Cramer homogène, donc l'unique solution est  $(0, 0, 0)$  :

$$\mathcal{S} = \{(0, 0, 0)\}.$$

### Exercice 5.20

Discuter et résoudre le système

$$(S_a) \quad \begin{cases} x + ay + (a-1)z = 0, \\ 3x + 2y + az = 3, \\ (a-1)x + ay + (a+1)z = a, \end{cases}$$

d'inconnue  $(x, y, z) \in \mathbb{R}^3$  et de paramètre  $a \in \mathbb{R}$ .

#### Correction

Comme le système contient un paramètre, on commence par calculer le déterminant de la matrice associée :

$$\begin{vmatrix} 1 & a & a-1 \\ 3 & 2 & a \\ a-1 & a & a+1 \end{vmatrix} = 2(a+1) + a^2(a-1) + 3a(a-1) - 2(a-1)^2 - a^2 - 3a(a+1) = a^2(a-4).$$

Le système est de Cramer si et seulement si ce déterminant est non nul, donc

$$(S_a) \text{ est de Cramer si et seulement } a \in \mathbb{R} \setminus \{0, 4\}.$$

Notons  $\mathcal{S}$  l'ensemble des solutions.

- ★ Étude du cas  $a = 0$ . Le système s'écrit

$$(S_0) \quad \begin{cases} x - z = 0, \\ 3x + 2y = 3, \\ -x + z = 0, \end{cases}$$

donc  $z = \kappa \in \mathbb{R}$ ,  $y = \frac{3-3\kappa}{2}$  et  $x = \kappa$ , ainsi

$$\mathcal{S} = \left\{ \left( \kappa, \frac{3-3\kappa}{2}, \kappa \right) \mid \kappa \in \mathbb{R} \right\}.$$

- ★ Étude du cas  $a = 4$ . Le système s'écrit

$$(S_4) \quad \begin{cases} x + 4y + 3z = 0, \\ 3x + 2y + 4z = 3, \\ 3x + 4y + 5z = 4, \end{cases}$$

On utilise la méthode du pivot de GAUSS :

$$\begin{cases} x + 4y + 3z = 0, \\ 3x + 2y + 4z = 3, \\ 3x + 4y + 5z = 4, \end{cases} \xrightarrow[L_2 \leftarrow L_2 - 3L_1]{L_3 \leftarrow L_3 - 3L_1} \begin{cases} x + 4y + 3z = 0, \\ -10y - 5z = 3, \\ -8y - 4z = 4, \end{cases} \xrightarrow[L_3 \leftarrow 10L_3 - 8L_2]{ } \begin{cases} x + 4y + 3z = 0, \\ -10y - 5z = 3, \\ 0 = 16. \end{cases}$$

La dernière équation est impossible donc

$$\mathcal{S} = \emptyset.$$

- ★ Étude du cas  $a \in \mathbb{R} \setminus \{-3, 0\}$ . On utilise la méthode du pivot de GAUSS :

$$\begin{cases} x + ay + (a-1)z = 0, \\ 3x + 2y + az = 3, \\ (a-1)x + ay + (a+1)z = a, \end{cases} \xrightarrow[L_3 \leftarrow L_3 - (a-1)L_1]{ } \begin{cases} x + ay + (a-1)z = 0, \\ (2-3a)y + (3-2a)z = 3, \\ (2-a)ay + (3-a)az = a, \end{cases}$$

$$\xrightarrow{L_3 \leftarrow L_3 - \frac{(2-a)a}{(2-3a)L_2} L_2} \begin{cases} x & +ay & +(a-1)z=0, \\ & (2-3a)y+(3-2a)z=3, \\ & -\frac{a^2(a-4)}{3a-2}z=\frac{4a}{3a-2}. \end{cases}$$

On obtient  $z = -\frac{4}{a(a-4)}$ ,  $y = -\frac{a-6}{a(a-4)}$ ,  $x = \frac{a^2-2a-4}{a(a-4)}$ , ainsi

$$\mathcal{S} = \left\{ \left( \frac{a^2-2a-4}{a(a-4)}, -\frac{a-6}{a(a-4)}, -\frac{4}{a(a-4)} \right) \right\}.$$

### Exercice 5.21

Calculer  $\mathbb{A}^{-1}$  où  $\mathbb{A}$  est la matrice  $\begin{pmatrix} 1 & 0 & -1 \\ 4 & -1 & -2 \\ -2 & 0 & 1 \end{pmatrix}$ .

**Correction**

$$\begin{aligned} [\mathbb{A}|\mathbb{I}_3] &= \left( \begin{array}{ccc|ccc} 1 & 0 & -1 & 1 & 0 & 0 \\ 4 & -1 & -2 & 0 & 1 & 0 \\ -2 & 0 & 1 & 0 & 0 & 1 \end{array} \right) \xrightarrow{\substack{L_1 \leftarrow L_1 \\ L_2 \leftarrow L_2 - 4L_1 \\ L_3 \leftarrow L_3 + 2L_1}} \left( \begin{array}{ccc|ccc} 1 & 0 & -1 & 1 & 0 & 0 \\ 0 & -1 & 2 & -4 & 1 & 0 \\ 0 & 0 & -1 & -2 & 0 & 1 \end{array} \right) \\ &\xrightarrow{\substack{L_1 \leftarrow L_1 \\ L_2 \leftarrow L_2 \\ L_3 \leftarrow L_3}} \left( \begin{array}{ccc|ccc} 1 & 0 & -1 & 1 & 0 & 0 \\ 0 & 1 & -2 & 4 & -1 & 0 \\ 0 & 0 & -1 & -2 & 0 & 1 \end{array} \right) \xrightarrow{\substack{L_1 \leftarrow L_1 - L_3 \\ L_2 \leftarrow L_2 - 2L_3 \\ L_3 \leftarrow -L_3}} \left( \begin{array}{ccc|ccc} 1 & 0 & 0 & -1 & 0 & -1 \\ 0 & 1 & 0 & 0 & -1 & -2 \\ 0 & 0 & 1 & -2 & 0 & -1 \end{array} \right) = [\mathbb{I}_3|\mathbb{A}^{-1}]. \end{aligned}$$

### Exercice 5.22

Calculer  $\mathbb{A}^{-1}$  où  $\mathbb{A}$  est la matrice  $\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 2 \\ 2 & 0 & 1 \end{pmatrix}$ .

**Correction**

$$\begin{aligned} [\mathbb{A}|\mathbb{I}_3] &= \left( \begin{array}{ccc|ccc} 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 2 & 0 & 1 & 0 \\ 2 & 0 & 1 & 0 & 0 & 1 \end{array} \right) \xrightarrow{\substack{L_1 \leftarrow L_1 \\ L_2 \leftarrow L_2 - 2L_1 \\ L_3 \leftarrow L_3 - 2L_1}} \left( \begin{array}{ccc|ccc} 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 2 & 0 & 1 & 0 \\ 0 & 0 & -1 & -2 & 0 & 1 \end{array} \right) \\ &\xrightarrow{\substack{L_1 \leftarrow L_1 \\ L_2 \leftarrow L_2 \\ L_3 \leftarrow L_3}} \left( \begin{array}{ccc|ccc} 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 2 & 0 & 1 & 0 \\ 0 & 0 & -1 & -2 & 0 & 1 \end{array} \right) \xrightarrow{\substack{L_1 \leftarrow L_1 + L_3 \\ L_2 \leftarrow L_2 + 2L_3 \\ L_3 \leftarrow -L_3}} \left( \begin{array}{ccc|ccc} 1 & 0 & 0 & -1 & 0 & 1 \\ 0 & 1 & 0 & 0 & -1 & 2 \\ 0 & 0 & 1 & 2 & 0 & -1 \end{array} \right) = [\mathbb{I}_3|\mathbb{A}^{-1}]. \end{aligned}$$

### Exercice 5.23

Calculer les inverses des matrices suivantes (si elles existent) :

$$\mathbb{A} = \begin{pmatrix} 2 & -3 \\ 4 & 5 \end{pmatrix}, \quad \mathbb{B} = \begin{pmatrix} 1 & 5 & -3 \\ 2 & 11 & 1 \\ 2 & 9 & -11 \end{pmatrix}, \quad \mathbb{C} = \begin{pmatrix} 1 & 5 & -3 \\ 2 & 11 & 1 \\ 1 & 4 & -10 \end{pmatrix}.$$

**Correction**

$\det(\mathbb{A}) = 22 \neq 0$  donc  $\mathbb{A}$  est inversible et on trouve

$$\mathbb{A}^{-1} = \frac{1}{22} \begin{pmatrix} 5 & 3 \\ -4 & 2 \end{pmatrix}.$$

$\det(\mathbb{B}) = 2 \neq 0$  donc  $\mathbb{B}$  est inversible et on trouve

$$\mathbb{B}^{-1} = \frac{1}{2} \begin{pmatrix} -130 & 28 & 38 \\ 24 & -5 & -7 \\ -4 & 1 & 1 \end{pmatrix}.$$

$\det(\mathbb{C}) = 0$  donc  $\mathbb{C}$  n'est pas inversible.

**Exercice 5.24**Soit  $\mathbb{A}$  la matrice

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

1. Calculer  $\det(\mathbb{A})$ .
2. Si  $\det(\mathbb{A}) \neq 0$ , calculer  $\mathbb{A}^{-1}$ .

**Correction**

1. Pour calculer le déterminant de la matrice  $\mathbb{A}$  on développe par rapport à la première ligne

$$\det(\mathbb{A}) = 1 \cdot \det(\mathbb{A}_{11}) - 0 \cdot \det(\mathbb{A}_{12}) + 0 \cdot \det(\mathbb{A}_{13}) - (-1) \cdot \det(\mathbb{A}_{14}) = \det\begin{pmatrix} 1 & -1 & -1 \\ 2 & -1 & -2 \\ 2 & 0 & -2 \end{pmatrix} + \det\begin{pmatrix} 1 & 1 & -1 \\ 1 & 2 & -1 \\ 1 & 2 & 0 \end{pmatrix}.$$

On note que la première colonne de la sous-matrice  $\mathbb{A}_{11}$  est l'opposée de la deuxième colonne, ainsi le déterminant de  $\mathbb{A}_{11}$  est nul et il ne reste plus qu'à calculer le déterminant de  $\mathbb{A}_{14}$  (par exemple en utilisant la règle de SARRUS).

$$\det(\mathbb{A}) = 0 + \det\begin{pmatrix} 1 & 1 & -1 \\ 1 & 2 & -1 \\ 1 & 2 & 0 \end{pmatrix} = 1.$$

2. Calculons  $\mathbb{A}^{-1}$  avec l'une des deux méthodes suivantes :

**Méthode de Gauss**

$$\begin{aligned} [\mathbb{A} | \mathbb{I}_4] &= \left( \begin{array}{cccc|cccccc} 1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 \\ 1 & 1 & -1 & -1 & 0 & 1 & 0 & 0 \\ 1 & 2 & -1 & -2 & 0 & 0 & 1 & 0 \\ 1 & 2 & 0 & -2 & 0 & 0 & 0 & 1 \end{array} \right) \xrightarrow[L_3 \leftarrow L_3 - L_1]{L_2 \leftarrow L_2 - L_1} \left( \begin{array}{cccc|cccccc} 1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & -1 & 1 & 0 & 0 \\ 0 & 2 & -1 & -1 & -1 & 0 & 1 & 0 \\ 0 & 2 & 0 & -1 & -1 & 0 & 0 & 1 \end{array} \right) \\ &\xrightarrow[L_3 \leftarrow L_3 - 2L_2]{L_4 \leftarrow L_4 - 2L_2} \left( \begin{array}{cccc|cccccc} 1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 1 & -2 & 1 & 0 \\ 0 & 0 & 2 & -1 & 1 & -2 & 0 & 1 \end{array} \right) \xrightarrow[L_2 \leftarrow L_2 + L_3]{L_4 \leftarrow L_4 - 2L_3} \left( \begin{array}{cccc|cccccc} 1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 1 & -1 & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 1 & -1 & 2 & -2 & 1 \end{array} \right) \\ &\xrightarrow[L_1 \leftarrow L_1 + L_4]{L_2 \leftarrow L_2 + L_4} \left( \begin{array}{cccc|cccccc} 1 & 0 & 0 & 0 & 0 & 2 & -2 & 1 \\ 0 & 1 & 0 & 0 & -1 & 1 & -1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 1 & -1 & 2 & -2 & 1 \end{array} \right) = [\mathbb{I}_4 | \mathbb{A}^{-1}]. \end{aligned}$$

**Méthode de Cramer**

- \* On calcule la matrice des cofacteurs des éléments de  $\mathbb{A}$ , appelée comatrice de  $\mathbb{A}$  :

$$\text{comatrice} = \begin{pmatrix} + \begin{vmatrix} 1 & -1 & -1 \\ 2 & -1 & -2 \\ 2 & 0 & -2 \end{vmatrix} & - \begin{vmatrix} 1 & -1 & -1 \\ 1 & -1 & -2 \\ 1 & 0 & -2 \end{vmatrix} & + \begin{vmatrix} 1 & 1 & -1 \\ 1 & 2 & -2 \\ 1 & 2 & -2 \end{vmatrix} & - \begin{vmatrix} 1 & 1 & -1 \\ 1 & 2 & -1 \\ 1 & 2 & 0 \end{vmatrix} \\ - \begin{vmatrix} 0 & 0 & -1 \\ 2 & -1 & -2 \\ 2 & 0 & -2 \end{vmatrix} & + \begin{vmatrix} 1 & 0 & -1 \\ 1 & -1 & -2 \\ 1 & 0 & -2 \end{vmatrix} & - \begin{vmatrix} 1 & 0 & -1 \\ 1 & 2 & -2 \\ 1 & 2 & -2 \end{vmatrix} & + \begin{vmatrix} 1 & 0 & 0 \\ 1 & 2 & -1 \\ 1 & 2 & 0 \end{vmatrix} \\ + \begin{vmatrix} 0 & 0 & -1 \\ 1 & -1 & -1 \\ 2 & 0 & -2 \end{vmatrix} & - \begin{vmatrix} 1 & 0 & -1 \\ 1 & -1 & -1 \\ 1 & 0 & -2 \end{vmatrix} & + \begin{vmatrix} 1 & 0 & -1 \\ 1 & 1 & -1 \\ 1 & 2 & -2 \end{vmatrix} & - \begin{vmatrix} 1 & 0 & 0 \\ 1 & 1 & -1 \\ 1 & 2 & 0 \end{vmatrix} \\ - \begin{vmatrix} 0 & 0 & -1 \\ 1 & -1 & -1 \\ 2 & -1 & -2 \end{vmatrix} & + \begin{vmatrix} 1 & 0 & -1 \\ 1 & -1 & -1 \\ 1 & -1 & -2 \end{vmatrix} & - \begin{vmatrix} 1 & 0 & -1 \\ 1 & 1 & -1 \\ 1 & 2 & -2 \end{vmatrix} & + \begin{vmatrix} 1 & 0 & 0 \\ 1 & 1 & -1 \\ 1 & 2 & -1 \end{vmatrix} \end{pmatrix} = \begin{pmatrix} 0 & -1 & 0 & -1 \\ 2 & 1 & 0 & 2 \\ -2 & -1 & -1 & -2 \\ 1 & 1 & 1 & 1 \end{pmatrix};$$

- \* on transpose la comatrice de  $\mathbb{A}$  :

$$\text{comatrice}^T = \begin{pmatrix} 0 & 2 & -2 & 1 \\ -1 & 1 & -1 & 1 \\ 0 & 0 & -1 & 1 \\ -1 & 2 & -2 & 1 \end{pmatrix};$$

\* on divise par  $\det(\mathbb{A})$  et on obtient

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

```
A=[1 0 0 -1; 1 1 -1 -1; 1 2 -1 -2; 1 2 0 -2]
det(A)
inv(A)
```

## Méthode de Gauss pour des systèmes rectangulaires (sur ou sous déterminés)

### Exercice 5.25

Vrai ou faux ?

- ① Un système linéaire de 4 équations à 3 inconnues dont les secondes membres sont nuls n'a que la solution nulle.
- ② Un système linéaire de 3 équations à 4 inconnues dont les secondes membres sont nuls a des solutions non nulles.

#### Correction

- ① Faux. Contrexemple : un système linéaire où toutes les équations sont identiques.
- ② Vrai :  $\text{rg}(\mathbb{A}) \leq 3$ ,  $\text{rg}([\mathbb{A}|\mathbf{b}]) \leq 3$ ; comme les secondes membres sont nuls alors  $\text{rg}([\mathbb{A}|\mathbf{b}]) = \text{rg}(\mathbb{A})$  donc il admet forcément des solutions ; comme il y a 4 inconnues, alors on a une infinité de solutions.

### Exercice 5.26

Résoudre le système

$$(S) \quad \begin{cases} -2x + y + z = 0, \\ x - 2y + z = 0, \end{cases}$$

d'inconnue  $(x, y, z) \in \mathbb{R}^3$ .

#### Correction

(S) est équivalent au système

$$\begin{cases} -2x + y + z = 0, \\ -3y + 3z = 0, \end{cases}$$

qui admet une infinité de solutions de la forme  $(\kappa, \kappa, \kappa)$  pour  $\kappa \in \mathbb{R}$ .

### Exercice 5.27

Soit le système linéaire

$$(S) \quad \begin{cases} x_1 + x_2 - 2x_3 + 4x_4 = 6, \\ -3x_1 - 3x_2 + 6x_3 - 12x_4 = b. \end{cases}$$

1. Pour quelle valeur de  $b$  le système est-il possible ?
2. Donner à  $b$  la valeur trouvée au point précédent et calculer la solution complète du système.

#### Correction

(S) est équivalent au système

$$\begin{cases} x_1 + x_2 - 2x_3 + 4x_4 = 6, \\ 0 = b + 18. \end{cases}$$

1. (S) est possible si et seulement si  $b = -18$ .
2. Si  $b = -18$ , (S) admet  $\infty^3$  solutions de la forme  $(x_1, x_2, x_3, x_4) = (6 - a + 2b - 4c, a, b, c)$  avec  $a, b, c \in \mathbb{R}$ .

### Exercice 5.28

Résoudre le système

$$(S) \quad \begin{cases} x + 2y + z = -1, \\ 2x + y - z = 1, \\ -x + y + 2z = -2, \\ x + y + z = 4. \end{cases}$$

**Correction**

(S) étant un système de 4 équations à 3 inconnues, on considère le sous-système carré d'ordre 3

$$(S') \quad \begin{cases} x+2y+z=-1, \\ 2x+y-z=1, \\ -x+y+2z=-2, \end{cases}$$

qu'on peut résoudre par la méthode du pivot de GAUSS

$$\begin{cases} x+2y+z=-1, \\ 2x+y-z=1, \\ -x+y+2z=-2, \end{cases} \xrightarrow[L_2-L_2-2L_1]{L_3-L_3+L_1} \begin{cases} x+2y+z=-1, \\ -3y-3z=3, \\ 3y+3z=-3, \end{cases} \xrightarrow[L_3+L_3+L_2]{ } \begin{cases} x+2y+z=-1, \\ -3y-3z=3, \\ 0=0, \end{cases}$$

qui admet une infinité de solutions de la forme  $(1+\kappa, -1-\kappa, \kappa)$  pour  $\kappa \in \mathbb{R}$ . Cherchons parmi ces solutions celles qui vérifient l'équation de (S) qui n'apparaît pas dans (S') : pour  $(x, y, z) = (1+\kappa, -1-\kappa, \kappa)$  on a  $x+y+z = 1+\kappa - 1 - \kappa + \kappa = \kappa$  donc  $x+y+z = 4$  si et seulement si  $\kappa = 4$  ainsi (S) admet l'unique solution  $(5, -5, 4)$ .

**Exercice 5.29**

Déterminer si le système suivant a une solution non nulle. Dans le cas affirmatif trouver la(les) solution(s) et expliquer pourquoi :

$$(S) \quad \begin{cases} x-2y+2z=0, \\ 2x+y-2z=0, \\ 3x+4y-6z=0, \\ 3x-11y+12z=0. \end{cases}$$

**Correction**

(S) étant un système de 4 équations à 3 inconnues, on considère le sous-système carré d'ordre 3

$$(S') \quad \begin{cases} x-2y+2z=0, \\ 2x+y-2z=0, \\ 3x+4y-6z=0, \end{cases}$$

qu'on peut résoudre par la méthode du pivot de GAUSS

$$\begin{cases} x-2y+2z=0, \\ 2x+y-2z=0, \\ 3x+4y-6z=0, \end{cases} \xrightarrow[L_2-L_2-2L_1]{L_3-L_3-3L_1} \begin{cases} x-2y+2z=0, \\ 5y-6z=0, \\ 10y-12z=0, \end{cases} \xrightarrow[L_3-L_3-2L_2]{ } \begin{cases} x-2y+2z=0, \\ 5y-6z=0, \\ 0=0, \end{cases}$$

qui admet une infinité de solutions de la forme  $(2\kappa, 6\kappa, 5\kappa)$  pour  $\kappa \in \mathbb{R}$ . Cherchons parmi ces solutions celles qui vérifient l'équation de (S) qui n'apparaît pas dans (S') : pour  $(x, y, z) = (2\kappa, 6\kappa, 5\kappa)$  on a  $3x-11y+12z = 6\kappa - 66\kappa + 60\kappa = 0$  donc  $3x-11y+12z = 0$  pour tout  $\kappa \in \mathbb{R}$  ainsi (S) admet une infinité de solutions de la forme  $(2\kappa, 6\kappa, 5\kappa)$  pour  $\kappa \in \mathbb{R}$ .

**Factorisation LU et systèmes linéaires carrés****Exercice 5.30**

Soit  $\mathbb{A}$  une matrice,  $\mathbb{A} \in \mathcal{M}_{n,n}(\mathbb{R})$ .

1. Rappeler les conditions nécessaires et suffisantes pour l'existence d'une factorisation LU de la matrice  $\mathbb{A}$  et préciser les définitions de L et U.
2. On suppose L et U construites (*i.e.* on dispose de tous les coefficients  $\ell_{i,j}$  et  $u_{i,j}$  de L et U), écrire l'algorithme de résolution de  $\mathbb{A}\mathbf{x} = \mathbf{b}$ , avec  $\mathbf{b} \in \mathcal{M}_{n,1}(\mathbb{R})$  donné.
3. Soit la matrice  $\mathbb{A}$  suivante :

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

Construire à la main les matrices L et U de la factorisation LU.

**Correction**

- Pour une matrice quelconque  $\mathbb{A} \in \mathcal{M}_{n,n}(\mathbb{R})$ , la factorisation  $\mathbb{L}\mathbb{U}$  (sans pivot) existe et est unique ssi les sous-matrices principales  $\mathbb{A}_i$  de  $\mathbb{A}$  d'ordre  $i = 1, \dots, n-1$  (celles que l'on obtient en restreignant  $\mathbb{A}$  à ses  $i$  premières lignes et colonnes) ne sont pas singulières (autrement dit si les mineurs principaux, i.e. les déterminants des sous-matrices principales, sont non nuls). On peut identifier des classes de matrices particulières pour lesquelles les hypothèses de cette proposition sont satisfaites. Mentionnons par exemple : les matrices à diagonale strictement dominante, les matrices réelles symétriques définies positives. Une technique qui permet d'effectuer la factorisation  $\mathbb{L}\mathbb{U}$  pour toute matrice  $\mathbb{A}$  inversible, même quand les hypothèses de cette proposition ne sont pas vérifiées, est la méthode du pivot par ligne : il suffit d'effectuer une permutation convenable des lignes de la matrice originale  $\mathbb{A}$  à chaque étape  $k$  où un terme diagonal  $a_{kk}$  s'annule.
- Une fois calculées les matrices  $\mathbb{L}$  et  $\mathbb{U}$ , résoudre le système linéaire  $\mathbb{A}\mathbf{x} = \mathbf{b}$ , avec  $\mathbf{b} \in \mathcal{M}_{n,1}(\mathbb{R})$  donné consiste simplement à résoudre successivement

2.1. le système triangulaire inférieur  $\mathbb{L}\mathbf{y} = \mathbf{b}$  par l'algorithme

$$y_1 = b_1, \quad y_i = b_i - \sum_{j=1}^{i-1} \ell_{ij} y_j, \quad i = 2, \dots, n$$

2.2. le système triangulaire supérieure  $\mathbb{U}\mathbf{x} = \mathbf{y}$  par l'algorithme

$$x_n = \frac{y_n}{u_{nn}}, \quad x_i = \frac{1}{u_{ii}} \left( y_i - \sum_{j=i+1}^n u_{ij} x_j \right), \quad j = n-1, \dots, 1$$

3. Factorisation :

$$\begin{pmatrix} 3 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{pmatrix} \xrightarrow{L_2 \leftarrow L_2 - \frac{-1}{3}L_1} \begin{pmatrix} 3 & -1 & -1 \\ 0 & \frac{8}{3} & -\frac{4}{3} \\ 0 & -\frac{4}{3} & \frac{8}{3} \end{pmatrix} \xrightarrow{L_3 \leftarrow L_3 - \frac{-4/3}{8/3}L_2} \begin{pmatrix} 3 & -1 & -1 \\ 0 & \frac{8}{3} & -\frac{4}{3} \\ 0 & 0 & 2 \end{pmatrix}.$$

Par conséquent

$$\mathbb{L} = \begin{pmatrix} 1 & 0 & 0 \\ -\frac{1}{3} & 1 & 0 \\ -\frac{1}{3} & -\frac{1}{2} & 1 \end{pmatrix} \quad \text{et} \quad \mathbb{U} = \begin{pmatrix} 3 & -1 & -1 \\ 0 & \frac{8}{3} & -\frac{4}{3} \\ 0 & 0 & 2 \end{pmatrix}.$$

**Exercice 5.31**

Résoudre les systèmes linéaires suivants :

$$\begin{cases} x - 5y - 7z = 3 \\ 2x - 13y - 18z = 3 \\ 3x - 27y - 36z = 3 \end{cases} \quad \text{et} \quad \begin{cases} x - 5y - 7z = 6 \\ 2x - 13y - 18z = 0 \\ 3x - 27y - 36z = -3 \end{cases} \quad \text{et} \quad \begin{cases} x - 5y - 7z = 0 \\ 2x - 13y - 18z = 3 \\ 3x - 27y - 36z = 6. \end{cases}$$

**Correction**

Le trois systèmes s'écrivent sous forme matricielle

$$\begin{pmatrix} 1 & -5 & -7 \\ 2 & -13 & -18 \\ 3 & -27 & -36 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \\ 3 \end{pmatrix} \quad \text{et} \quad \begin{pmatrix} 1 & -5 & -7 \\ 2 & -13 & -18 \\ 3 & -27 & -36 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 6 \\ 0 \\ -3 \end{pmatrix} \quad \text{et} \quad \begin{pmatrix} 1 & -5 & -7 \\ 2 & -13 & -18 \\ 3 & -27 & -36 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 3 \\ 6 \end{pmatrix}$$

On remarque que seul le terme source change. On calcul d'abord la décomposition  $\mathbb{L}\mathbb{U}$  de la matrice  $\mathbb{A}$  :

$$\begin{pmatrix} 1 & -5 & -7 \\ 2 & -13 & -18 \\ 3 & -27 & -36 \end{pmatrix} \xrightarrow{L_2 \leftarrow L_2 - 2L_1} \begin{pmatrix} 1 & -5 & -7 \\ 0 & -3 & -4 \\ 0 & -12 & -15 \end{pmatrix} \xrightarrow{L_3 \leftarrow L_3 - 4L_2} \begin{pmatrix} 1 & -5 & -7 \\ 0 & -3 & -4 \\ 0 & 0 & 1 \end{pmatrix}$$

donc

$$\mathbb{L} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 4 & 1 \end{pmatrix} \quad \mathbb{U} = \begin{pmatrix} 1 & -5 & -7 \\ 0 & -3 & -4 \\ 0 & 0 & 1 \end{pmatrix}$$

Pour résoudre chaque système linéaire on résout les systèmes triangulaires  $\mathbb{L}\mathbf{y} = \mathbf{b}$  et  $\mathbb{U}\mathbf{x} = \mathbf{y}$ .

1. Pour le premier système on a

$$\begin{aligned} \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 4 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} &= \begin{pmatrix} 3 \\ 3 \\ 3 \end{pmatrix} \implies y_1 = 3, \quad y_2 = -3, \quad y_3 = 6; \\ \begin{pmatrix} 1 & -5 & -7 \\ 0 & -3 & -4 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} &= \begin{pmatrix} 3 \\ -3 \\ 6 \end{pmatrix} \implies x_3 = 6, \quad x_2 = -7, \quad x_1 = 10. \end{aligned}$$

2. Pour le seconde système on a

$$\begin{aligned} \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 4 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} &= \begin{pmatrix} 6 \\ 0 \\ -3 \end{pmatrix} \implies y_1 = 6, \quad y_2 = -12, \quad y_3 = 27; \\ \begin{pmatrix} 1 & -5 & -7 \\ 0 & -3 & -4 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} &= \begin{pmatrix} 6 \\ -12 \\ 27 \end{pmatrix} \implies x_3 = 27, \quad x_2 = -32, \quad x_1 = 35. \end{aligned}$$

3. Pour le dernier système on a

$$\begin{aligned} \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 4 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} &= \begin{pmatrix} 0 \\ 3 \\ 6 \end{pmatrix} \implies y_1 = 0, \quad y_2 = 3, \quad y_3 = -6; \\ \begin{pmatrix} 1 & -5 & -7 \\ 0 & -3 & -4 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} &= \begin{pmatrix} 6 \\ -12 \\ 27 \end{pmatrix} \implies x_3 = -6, \quad x_2 = 7, \quad x_1 = -7. \end{aligned}$$

### ★ Exercice 5.32

1. Implémenter une fonction appelée `descente` permettant de résoudre un système linéaire dont la matrice est triangulaire inférieure. La syntaxe doit être `function y=descente(L,b)` où **b** est un vecteur colonne de  $\mathbb{R}^n$  et **L** une matrice de  $\mathbb{R}^{n \times n}$  triangulaire inférieure. On doit obtenir un vecteur colonne de  $\mathbb{R}^n$  solution du système linaire  $\mathbf{L}\mathbf{y} = \mathbf{b}$ . Écrire un script appelé `TESTdescente.m` pour tester cette fonction sur l'exemple suivant : pour

$$\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 3 & 0 \\ 4 & 5 & 6 \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} 1 \\ 8 \\ 32 \end{pmatrix}$$

on doit obtenir

$$\mathbf{y} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$

Dans ce script on pourra aussi comparer les résultats obtenus par notre fonction `descente` et la commande d'Octave `L\b` sur différents systèmes linéaires triangulaires.

2. Implémenter une fonction appelée `remontee` permettant de résoudre un système linéaire dont la matrice est triangulaire supérieure. La syntaxe doit être `x=remontee(U,y)` où **y** est un vecteur colonne de  $\mathbb{R}^n$  et **U** une matrice de  $\mathbb{R}^{n \times n}$  triangulaire supérieure. On doit obtenir un vecteur colonne de  $\mathbb{R}^n$  solution du système linaire  $\mathbf{Ux} = \mathbf{y}$ . Écrire un script appelé `TESTremontee.m` pour tester cette fonction sur l'exemple suivant : pour

$$\mathbf{U} = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 6 \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} 14 \\ 23 \\ 18 \end{pmatrix}$$

on doit obtenir

$$\mathbf{x} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$

Dans ce script on pourra aussi comparer les résultats obtenus par notre fonction `remontee` et la commande d'Octave

U\y sur différents systèmes linéaires triangulaires.

3. Implémenter une fonction appelée `mylu` permettant de calculer la factorisation LU d'une matrice A par la méthode de GAUSS. La syntaxe doit être `[L, U] = mylu(A)` où A est une matrice de  $\mathbb{R}^{n \times n}$  non singulière et b est un vecteur colonne de  $\mathbb{R}^n$ . On doit obtenir L et U deux matrices de  $\mathbb{R}^{n \times n}$  triangulaires inférieur et supérieur respectivement telles que LU = A. Écrire un script appelé `TESTmylu.m` pour tester cette fonction sur l'exemple suivant : pour

$$A = \begin{pmatrix} 1 & 0 & 3 \\ 2 & 2 & 2 \\ 3 & 6 & 4 \end{pmatrix}$$

on doit obtenir

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

Dans ce script on pourra aussi comparer les résultats obtenus par notre fonction `mylu` et la fonction d'Octave `lu` sur différentes matrices.

4. Écrire une fonction appelé `syslin` permettant de résoudre le système linéaire  $Ax = b$  en utilisant la factorisation LU de la matrice A puis la résolution des systèmes linéaires  $Ly = b$  et  $Ux = y$ . La syntaxe doit être `[x] = syslin (A, b)` où A est une matrice de  $\mathbb{R}^{n \times n}$  non singulière. On doit obtenir x un vecteur colonne de  $\mathbb{R}^n$  solution du système linéaire  $Ax = b$ . Écrire un script appelé `TESTsyslin.m` pour tester cette fonction sur l'exemple suivant : pour

$$A = \begin{pmatrix} 1 & 0 & 3 \\ 2 & 2 & 2 \\ 3 & 6 & 4 \end{pmatrix} \quad b = \begin{pmatrix} 4 \\ 6 \\ 13 \end{pmatrix}$$

on doit obtenir

$$x = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

Dans ce script on pourra aussi comparer les résultats obtenus par notre fonction `syslin` et la commande d'Octave `A\b` sur différents systèmes linéaires.

5. Écrire une fonction appelée `mydet` permettant de calculer le déterminant d'une matrice A en utilisant la factorisation LU de la matrice A. La syntaxe doit être `[d] = mydet (A)` où A est une matrice de  $\mathbb{R}^{n \times n}$  non singulière. On doit obtenir  $d = \det(A)$ . Écrire un script appelé `TESTmydet.m` pour tester cette fonction sur l'exemple suivant : pour

$$\det \begin{pmatrix} 1 & 0 & 3 \\ 2 & 2 & 2 \\ 3 & 6 & 4 \end{pmatrix} = 14$$

Dans ce script on pourra aussi comparer les résultats obtenus par notre fonction `mydet` et la commande d'Octave `det(A)` sur différentes matrices.

6. Écrire une fonction appelée `myinv` permettant de calculer la matrice  $A^{-1}$  d'une matrice A en utilisant la factorisation LU de la matrice A et la résolution des  $2n$  systèmes linéaires  $Ly = e_j$  et  $Ux = y$  avec  $e_j$  le vecteur  $(e_j)_i = \delta_{ij}$ . La syntaxe doit être `[invA] = myinv(A)` où A est une matrice de  $\mathbb{R}^{n \times n}$  non singulière. On doit obtenir  $A^{-1}$  une matrice de  $\mathbb{R}^{n \times n}$  telle que  $A^{-1}A = AA^{-1} = I_n$ . Écrire un script appelé `TESTmyinv.m` pour tester cette fonction sur l'exemple suivant : pour

$$A = \begin{pmatrix} 1 & 0 & 3 \\ 2 & 2 & 2 \\ 3 & 6 & 4 \end{pmatrix}$$

on doit obtenir

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

Dans ce script on pourra aussi comparer les résultats obtenus par notre fonction `myinv` et la commande d'Octave

`inv(A)` sur différentes matrices.

7. Implémenter une fonction appelée `mylupivot` permettant de calculer la factorisation LU d'une matrice A par la méthode de GAUSS avec pivot, i.e.  $\mathbb{P}A = LU$ . La syntaxe doit être `[L,U,P]=mylupivot(A)` où A est une matrice de  $\mathbb{R}^{n \times n}$  non singulière.

Expliquer pourquoi on ne peut pas effectuer la factorisation LU de la matrice

$$A = \begin{pmatrix} 1 & 1 & 3 \\ 2 & 2 & 2 \\ 3 & 6 & 4 \end{pmatrix}$$

mais on peut effectuer la factorisation avec pivot. Calculer cette factorisation.

8. Implémenter une fonction appelée `syslinpivot.m` permettant de résoudre le système linéaire  $Ax = b$  en utilisant la factorisation LU avec pivot de la matrice A puis la résolution des systèmes linéaires  $Ly = Pb$  et  $Ux = y$ . La syntaxe doit être `[x]=syslinpivot(A,b)` où A est une matrice de  $\mathbb{R}^{n \times n}$  non singulière. On doit obtenir x un vecteur colonne de  $\mathbb{R}^n$  solution du système linaire  $Ax = b$ . Écrire un script appelé `TESTsyslinpivot.m` pour tester cette fonction sur l'exemple suivant : pour

$$A = \begin{pmatrix} 1 & 1 & 3 \\ 2 & 2 & 2 \\ 3 & 6 & 4 \end{pmatrix} \quad b = \begin{pmatrix} 5 \\ 6 \\ 13 \end{pmatrix}$$

on doit obtenir

$$x = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

Dans ce script on pourra aussi comparer les résultats obtenus par notre fonction `syslinpivot` et la commande d'Octave `A\b` sur différents systèmes linéaires.

### Correction

1. Dans le fichier `descente.m` on écrit

```
function y=descente(L,b)
y(1)=b(1)/L(1,1);
for i=2:length(b)
    y(i)=(b(i)-dot(L(i,1:i-1),y(1:i-1)))/L(i,i);
end
```

et on teste cette fonction par exemple comme suit

```
L=[1 0 0; 2 3 0; 4 5 6]
b=[1; 8; 32]
y=descente(L,b),
% Pour vérifier notre résultat on peut
% soit comparer le résultat avec celui d'Octave
yOctave=L\b
% soit vérifier que Ly=b
Ly=L*y
```

2. Dans le fichier `remontee.m` on écrit

```
function x=remontee(U,y)
n=length(y);
x(n)=y(n)/U(n,n);
for i=n-1:-1:1
    x(i)=(y(i)-dot(U(i,i+1:n),x(i+1:n)))/U(i,i);
end
```

et on teste cette fonction par exemple comme suit

```
U=[1 2 3; 0 4 5; 0 0 6]
y=[14; 23; 18]
x=remontee(U,y),
% Pour vérifier notre résultat on peut
% soit comparer le résultat avec celui d'Octave
xOctave=U\b
% soit vérifier que Ux=y
Ux=U*x
```

3. Dans le fichier `mylu.m` on écrit

```
function [L,U]=mylu(A)
% Factorisation de Doolittle, i.e. L(i,i)=1
[n,m]=size(A);
if n ~= m
    error('A is not a square matrix');
else
    tol=1.0e-9;
```

```

for k=1:n-1
    for i=k+1:n
        if abs(A(k,k))<tol
            error("Utiliser pivot");
        else
            A(i,k)=A(i,k)/A(k,k);
            A(i,k+1:n)=A(i,k+1:n)-A(i,k)*A(k,k+1:n);
        end
    end
end
U=triu(A);
L=tril(A,-1)+eye(n);
end
end

```

et on teste cette fonction par exemple comme suit

```

printf(...
"=====\\n\
Test 1\\n\
=====\\n");
A=[1 0 3; 2 2 2; 3 6 4]
[L,U]=mylu(A)
% Verifions notre resultat i.e. LU=A
LU=L*U

printf(...
"=====\\n\
Test 2\\n\
=====\\n");
A=[1 2 3 4; 2 3 4 1; 3 4 1 2; 4 1 2 3]
[L,U]=mylu(A)
LU=L*U

printf(...
"=====\\n\
Test 3\\n\
=====\\n");
A=[1 1 3; 2 2 2; 3 6 4]
[L,U]=mylu(A)

```

4. Dans le fichier `syslin.m` on écrit

```

function x=syslin(A,b)
    [L,U]=mylu(A);
    y=descente(L,b)';
    x=remontee(U,y)';
end

```

et on teste cette fonction par exemple comme suit

```

A=[1 0 3; 2 2 2; 3 6 4]
b=[4; 6; 13]
x=syslin(A,b)
% Pour verifier notre resultat on peut
% comparer au resultat d'Octave
xOctave=A\b
% ou verifier que Ax=b
Ax=A*x

```

5. Dans le fichier `mydet.m` on écrit

```

function d=mydet(A)
    [L,U]=mylu(A);
    d=prod(diag(U));
end

```

et on teste cette fonction par exemple comme suit

```

A=[1 0 3; 2 2 2; 3 6 4]
d=mydet(A)
% Pour verifier notre resultat on peut
% comparer au resultat d'Octave
dOctave=det(A)

```

6. Dans le fichier `myinv.m` on écrit

```

function invA=myinv(A)
[n,m]=size(A);

```

```

for j=1:n
    b=zeros(n);
    b(j)=1;
    invA(:,j)=syslin(A,b);

```

```
end
end
```

et on teste cette fonction par exemple comme suit

```
A=[1 0 3; 2 2 2; 3 6 4]
invA=myinv(A)
% Pour verifier notre resultat on peut
% comparer au resultat d'Octave
invAoctave=inv(A)
% ou verifier que invA*A=A*invA=Identite
invA*A
A*invA
```

7. Dans le fichier mylupivot.m on écrit

```
function [L,U,P]=mylupivot(A)
% Factorisation de Doolittle, i.e. L(i,i)=1
[n,m]=size(A);
if n ~= m
    error('A non carree');
else
    tol=1.0e-9;
    P = eye(n);
    for k=1:n-1
        [maxVal ipiv] = max(abs(A(k:n,k)));
        % echange L(k) <-> L(i)
        ipiv+=k-1; % car ipiv demarre de k
        A([k ipiv],:)=A([ipiv k],:);
        P([k ipiv],:)=P([ipiv k],:);
        for i=k+1:n
            A(i,k)=A(i,k)/A(k,k);
            A(i,k+1:n)=A(i,k+1:n)-A(i,k)*A(k,k+1:n);
        end
    end
    U=triu(A);
    L=tril(A,-1)+eye(n);
end
end
```

et on teste cette fonction par exemple comme suit

```
printf...
=====
Test 1\n\
=====
% mylu et mylupivot donnent le meme resultat
A=[1 0 3; 2 2 2; 3 6 4]
[L,U,P]=mylupivot(A)
% Verifions notre resultat, i.e. LU=PA
LU=L*U
PA=P*A
% Comparons le resultat avec celui d'Octave
[Loctave,Uoctave,Poctave]=lu(A)

printf...
=====
Test 2\n\
=====
A=[1 1 3; 2 2 2; 3 6 4]
[L,U,P]=mylupivot(A)
% Verifions notre resultat, i.e. LU=PA
LU=L*U
PA=P*A
% Comparons le resultat avec celui d'Octave
[Loctave,Uoctave,Poctave]=lu(A)
% On ne peut pas utiliser mylu mais forcement mylupivot
```

```
% [L,U]=mylu(A)

printf(...
"=====\\n\
Test 3\\n\
=====\\n");
A=[1 2; 1 2]
% det(A)=0 mais on peut écrire A=LU car det(A_1)~0
[L,U,P]=mylupivot(A)
LU=L*U
PA=P*A

printf(...
"=====\\n\
Test 4\\n\
=====\\n");
A=[0 1; 1 0]
% det(A)~0 mais det(A_1)=0. On effectue alors le pivot
[L,U,P]=mylupivot(A)
LU=L*U
PA=P*A
```

On ne peut pas utiliser la factorisation LU sans pivot car  $\det(\mathbb{A}_1) = 1$  mais  $\det(\mathbb{A}_2) = \begin{pmatrix} 1 & 1 \\ 2 & 2 \end{pmatrix} = 0$ . Cependant on peut calculer la factorisation LU avec pivot car la matrice n'est pas singulière.

8. Dans le fichier syslinpivot.m on écrit

et on teste cette fonction par exemple comme suit

```
A=[1 0 3; 2 2 2; 3 6 4]
b=[4; 6; 13]
x=syslinpivot(A,b)
% Comparons le résultat à celui d'Octave
xOctave=A\b
% Vérifions que Ax=b
Ax=A*x

A=[1 1 3; 2 2 2; 3 6 4]
b=[5; 6; 13]
x=syslinpivot(A,b)
% Comparons le résultat à celui d'Octave
xOctave=A\b
% Vérifions que Ax=b
Ax=A*x
```

### ★ Exercice 5.33 (Matrices tridiagonales : algorithme de Thomas)

On considère la matrice tridiagonale inversible  $\mathbb{A} \in \mathbb{R}^{n \times n}$

$$\mathbb{A} = \begin{pmatrix} a_1 & c_1 & 0 & \dots & \dots & 0 \\ b_2 & a_2 & c_2 & \ddots & & \vdots \\ 0 & b_3 & a_3 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & b_{n-1} & a_{n-1} & c_{n-1} \\ 0 & \dots & \dots & 0 & b_n & a_n \end{pmatrix}$$

1. Montrer que les matrices  $\mathbb{L}$  et  $\mathbb{U}$  de la factorisation LU de  $\mathbb{A}$  sont bidiagonales, i.e. si  $a_{ij} = 0$  pour  $|i - j| > 1$  alors  $\ell_{ij} = 0$  pour  $i > 1 + j$  (et pour  $i < j$  car triangulaire inférieure) et  $u_{ij} = 0$  pour  $i < j - 1$  (et pour  $i > j$  car triangulaire supérieure).
2. On a montré au point précédent que les matrices  $\mathbb{L}$  et  $\mathbb{U}$  de la factorisation LU de  $\mathbb{A}$  sont bidiagonales, écrivons-les

sous la forme

$$\mathbb{L} = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & 0 \\ \beta_2 & 1 & \ddots & & & \vdots \\ 0 & \beta_3 & 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \beta_{n-1} & 1 & 0 \\ 0 & \dots & \dots & 0 & \beta_n & 1 \end{pmatrix}, \quad \mathbb{U} = \begin{pmatrix} \alpha_1 & \gamma_1 & 0 & \dots & \dots & 0 \\ 0 & \alpha_2 & \gamma_2 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ \vdots & & & \ddots & \alpha_{n-1} & \gamma_{n-1} \\ 0 & \dots & \dots & \dots & 0 & \alpha_n \end{pmatrix}.$$

Calculer  $(\alpha_1, \alpha_2, \dots, \alpha_n)$ ,  $(\beta_2, \beta_3, \dots, \beta_n)$  et  $(\gamma_1, \gamma_2, \dots, \gamma_{n-1})$  en fonction de  $(a_1, a_2, \dots, a_n)$ ,  $(b_2, b_3, \dots, b_n)$  et  $(c_1, c_2, \dots, c_{n-1})$ . En déduire un algorithme de factorisation.

3. À l'aide des formules trouvées au point précédent, écrire l'algorithme pour résoudre le système linéaire  $\mathbb{A}\mathbf{x} = \mathbf{f}$  où  $\mathbf{f} = (f_1, f_2, \dots, f_n)^T \in \mathbb{R}^n$ .

### Correction

1. Soit  $\mathbb{A}^{(k)}$ ,  $k = 0, \dots, n-1$  la matrice obtenue à l'étape  $k$  de la méthode de GAUSS, avec  $\mathbb{A}^{(0)} = \mathbb{A}$  et  $\mathbb{A}^{(n-1)} = \mathbb{U}$ . On montrera par récurrence sur  $k$  que  $\mathbb{A}^{(k)}$  est tridiagonale, i.e.  $a_{ij}^{(k)} = 0$  pour  $|i - j| > 1$ .

**Initialisation :** pour  $k = 0$ ,  $\mathbb{A}^{(0)} = \mathbb{A}$  qui est une matrice tridiagonale.

**Héritérité :** soit  $\mathbb{A}^{(k)}$  une matrice tridiagonale (i.e.  $a_{ij}^{(k)} = 0$  pour  $|i - j| > 1$ ) et montrons que  $\mathbb{A}^{(k+1)}$  l'est aussi.

- \* Si  $i \leq k$  alors  $a_{ij}^{(k+1)} = a_{ij}^{(k)} = 0$  (les lignes  $L_1, \dots, L_k$  de la matrice  $\mathbb{A}^{(k)}$  ne sont pas modifiées à l'étape  $k$ ).
- \* Soit  $i > k$ , alors les lignes  $L_{k+1}, \dots, L_n$  de la matrice  $\mathbb{A}^{(k)}$  vont être modifiées selon la relation

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} a_{kj}^{(k)}.$$

Pour chaque ligne  $i > k$ , considérons séparément les colonnes  $j \leq k$  et les colonnes  $j > k$ :

- \* si  $j \leq k$ ,  $a_{ij}^{(k+1)} = 0$  (zéros qu'on fait apparaître avec la méthode de GAUSS pour une matrice quelconque),
- \* soit  $j > k$ :
  - \* si  $j < i-1$ , comme  $i, j > k$  alors  $a_{ij}^{(k)} = 0$  et  $i > j+1 > k+1$ , c'est-à-dire  $i-k > 1$  et donc  $a_{ik}^{(k)} = 0$  et  $\ell_{ik}^{(k)} = 0$ . Donc  $a_{ij}^{(k+1)} = 0$ .
  - \* si  $j > i+1$ , comme  $i, j > k$  alors  $a_{ij}^{(k)} = 0$  et  $j > i+1 > k+1$ , c'est-à-dire  $j-k > 1$  et donc  $a_{kj}^{(k)} = 0$ . Donc  $a_{ij}^{(k+1)} = 0$ .

2. Les coefficients  $(\alpha_1, \alpha_2, \dots, \alpha_n)$ ,  $(\beta_2, \beta_3, \dots, \beta_n)$  et  $(\gamma_1, \gamma_2, \dots, \gamma_{n-1})$  se calculent en imposant l'égalité  $\mathbb{L}\mathbb{U} = \mathbb{A}$ . L'algorithme se déduit en parcourant les étapes de la méthode de GAUSS :

$$\mathbb{A}^{(0)} = \begin{pmatrix} a_1 & c_1 & 0 & \dots & \dots & 0 \\ b_2 & a_2 & c_2 & \ddots & & \vdots \\ 0 & b_3 & a_3 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & b_{n-1} & a_{n-1} & c_{n-1} \\ 0 & \dots & \dots & 0 & b_n & a_n \end{pmatrix}$$

$$\xrightarrow{\substack{L_2 \leftarrow L_2 - \beta_2 L_1 \\ \beta_2 = \frac{b_2}{a_1}}} \mathbb{A}^{(1)} = \begin{pmatrix} \alpha_1 = a_1 & \gamma_1 = c_1 & 0 & \dots & \dots & 0 \\ 0 & a_2 = a_2 - \beta_2 c_1 & \gamma_2 = c_2 & \ddots & & \vdots \\ 0 & b_3 & a_3 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & \ddots & b_{n-1} & a_{n-1} & c_{n-1} \\ 0 & \dots & \dots & 0 & b_n & a_n \end{pmatrix}$$

$$\frac{L_3 \leftarrow L_3 - \beta_3 L_2}{\beta_3 = \frac{b_3}{\alpha_2}} \mathbb{A}^{(2)} = \begin{pmatrix} \alpha_1 = a_1 & \gamma_1 = c_1 & 0 & \dots & \dots & 0 \\ 0 & \alpha_2 = a_2 - \beta_2 c_1 & \gamma_2 = c_2 & \ddots & & \vdots \\ 0 & 0 & \alpha_3 = a_3 - \beta_3 c_2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & b_{n-1} & a_{n-1} & c_{n-1} \\ 0 & \dots & \dots & 0 & b_n & a_n \end{pmatrix} \xrightarrow[L_4 \leftarrow L_4 - \beta_4 L_3]{\beta_4 = \frac{b_4}{\alpha_3}} [\dots]$$

$$[\dots] \frac{L_n \leftarrow L_n - \beta_n L_{n-1}}{\beta_n = \frac{b_n}{\alpha_n}} \mathbb{A}^{(n-1)} = \begin{pmatrix} \alpha_1 = a_1 & \gamma_1 = c_1 & 0 & \dots & \dots & 0 \\ 0 & \alpha_2 = a_2 - \beta_2 c_1 & \gamma_2 = c_2 & \ddots & & \vdots \\ 0 & 0 & \alpha_3 = a_3 - \beta_3 c_2 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & 0 & \alpha_{n-1} = a_{n-1} - \beta_{n-1} c_{n-2} & \gamma_{n-1} = c_{n-1} \\ 0 & \dots & \dots & 0 & 0 & \alpha_n = a_n - \beta_n c_{n-1} \end{pmatrix}$$

Donc  $\gamma_i = c_i$  pour  $i = 1, \dots, n$ ,  $\alpha_1 = a_1$  et on définit par récurrence

$$\begin{cases} \beta_i = \frac{b_i}{\alpha_{i-1}} \\ \alpha_i = a_i - \beta_i c_{i-1} \end{cases} \quad \text{pour } i = 2, \dots, n.$$

3. La résolution du système linéaire  $\mathbb{A}\mathbf{x} = \mathbf{f}$  se ramène à la résolution des deux systèmes linéaires  $\mathbb{L}\mathbf{y} = \mathbf{f}$  et  $\mathbb{U}\mathbf{x} = \mathbf{y}$ , pour lesquels on obtient les formules suivantes :

$$\begin{cases} y_1 = f_1, \\ y_i = f_i - \beta_i y_{i-1}, \quad \text{pour } i = 2, \dots, n, \end{cases}$$

$$\begin{cases} x_n = \frac{y_n}{\alpha_n}, \\ x_i = \frac{y_i - \gamma_i x_{i+1}}{\alpha_i}, \quad \text{pour } i = n-1, \dots, 1, \end{cases}$$

$$\text{i.e. } \begin{cases} y_1 = f_1, \\ y_i = f_i - \frac{b_i}{\alpha_{i-1} - \beta_i c_{i-1}} y_{i-1}, \quad \text{pour } i = 2, \dots, n; \end{cases}$$

$$\text{i.e. } \begin{cases} x_n = \frac{y_n}{\alpha_n}, \\ x_i = \frac{y_i - \gamma_i x_{i+1}}{\alpha_i - \beta_i c_{i-1}}, \quad \text{pour } i = n-1, \dots, 2, \\ x_1 = \frac{y_1 - \gamma_1 x_2}{\alpha_1}. \end{cases}$$

Dans le fichier `syslinThomas.m` on écrit

et on teste cette fonction par exemple comme suit

```
function x=syslinThomas(a,b,c,f)
    [alpha,beta]=myluThomas(a,b,c);
    y=descenteThomas(beta,f)';
    x=remonteeThomas(alpha,c,y)';
end
```

```
clear all;
```

```
b=ones(10,1);
a=2*b;
c=3*b;
A=spdiags([b,a,c],-1:1,10,10);
f=A*b;

xOctave=A\f
x=syslinThomas(a,b,c,f)
```

Ce fichier utilise les fonctions suivantes : Fichier `myluThomas.m` :

```
function [alpha,beta]=myluThomas(a,b,c)
% Factorisation de Doolittle, i.e. L(i,i)=1
% A tridiagonale
% Algorithme de Thomas
% a=[a(1),a(2),...,a(n-1),a(n)] idem alpha
% b=[0 ,b(2),...,b(n-1),b(n)] idem beta
% c=[c(1),c(2),...,c(n-1), 0 ] = gamma

n=length(a); %length(b)=length(c)

% Factorisation LU
alpha(1)=a(1);
for i=2:n
    beta(i)=b(i)/alpha(i-1);
```

```

alpha(i)=a(i)-beta(i)*c(i-1);
end
% L=diag(beta(2:n),-1)+eye(1)
% U=diag(c(1:n-1),1)+diag(alpha)
end

```

Fichier `descenteThomas.m`:

```

function y=descenteThomas(beta,f)
n=length(beta);
% Resolution Ly=f
% L=diag(beta(2:n),-1)+eye(1)
y(1)=f(1);
for i=2:n
    y(i)=f(i)-beta(i)*y(i-1);
end
end

```

Fichier `remonteeThomas.m`:

```

function x=remonteeThomas(alpha,c,y)
n=length(y);
% Resolution Ux=y
% U=diag(c(1:n-1),1)+diag(alpha)
x(n)=y(n)/alpha(n);
for i=n-1:-1:1
    x(i)=(y(i)-c(i)*x(i+1))/alpha(i);
end
end

```

### Exercice 5.34

Soit la matrice  $\mathbb{A} \in \mathbb{R}^{n \times n}$ ,  $n \geq 3$ , dont les éléments vérifient

- $a_{ij} = 1$  si  $i = j$  ou  $i = n$ ,
- $a_{ij} = -1$  si  $i < j$ ,
- $a_{ij} = 0$  sinon.

Calculer la factorisation  $\mathbb{L}\mathbb{U}$  de  $\mathbb{A}$ .

### Correction

Factorisation  $\mathbb{L}\mathbb{U}$  de la matrice  $\mathbb{A}$ :

$$\begin{array}{c}
\left( \begin{array}{cccccc} 1 & -1 & \dots & \dots & \dots & -1 \\ 0 & 1 & \ddots & & & \vdots \\ \vdots & \ddots & 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ 0 & \dots & \dots & 0 & 1 & -1 \\ 1 & 1 & 1 & \dots & 1 & 1 \end{array} \right) \xrightarrow{L_n \leftarrow L_n - \frac{1}{1}L_1} \left( \begin{array}{cccccc} 1 & -1 & \dots & \dots & \dots & -1 \\ 0 & 1 & \ddots & & & \vdots \\ \vdots & \ddots & 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ 0 & \dots & \dots & 0 & 1 & -1 \\ 0 & 2 & 2 & \dots & 2 & 2 \end{array} \right) \xrightarrow{L_n \leftarrow L_n - \frac{2}{1}L_2} \left( \begin{array}{cccccc} 1 & -1 & \dots & \dots & \dots & -1 \\ 0 & 1 & \ddots & & & \vdots \\ \vdots & \ddots & 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ 0 & \dots & \dots & 0 & 1 & -1 \\ 0 & 0 & 4 & \dots & 4 & 4 \end{array} \right) \\
[...] \xrightarrow{L_n \leftarrow L_n - \frac{2^{n-2}}{1}L_{n-1}} \left( \begin{array}{cccccc} 1 & -1 & \dots & \dots & \dots & -1 \\ 0 & 1 & \ddots & & & \vdots \\ \vdots & \ddots & 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ 0 & \dots & \dots & 0 & 1 & -1 \\ 0 & 0 & 0 & \dots & 0 & 2^{n-1} \end{array} \right).
\end{array}$$

On obtient les matrices

$$\mathbb{L} = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & 0 \\ 0 & 1 & \ddots & & & \vdots \\ 0 & 0 & 1 & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 1 & 0 \\ 1 & 2 & 4 & \dots & 2^{n-2} & 1 \end{pmatrix} \quad \text{et} \quad \mathbb{U} = \begin{pmatrix} 1 & -1 & \dots & \dots & \dots & -1 \\ 0 & 1 & \ddots & & & \vdots \\ 0 & 0 & 1 & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & 0 & \dots & 0 & 1 & -1 \\ 0 & 0 & 0 & \dots & 0 & 2^{n-1} \end{pmatrix}.$$

c'est-à-dire

- $\ell_{ii} = 1$  pour  $i = 1, \dots, n$ ,
  - $\ell_{ij} = 0$  si  $i < n$  et  $i \neq j$ ,
  - $\ell_{nj} = 2^{j-1}$  si  $j < n$ ;
  - $u_{ij} = a_{ij}$  pour  $i=1, \dots, n-1, j=1, \dots, n$ ,
  - $u_{nj} = 0$  si  $j < n$ ,
  - $u_{nn} = 2^{n-1}$ .

### Exercice 5.35

Considérons une matrice  $A \in \mathbb{R}^{n \times n}$  (avec  $n \geq 3$ ) dont les éléments vérifient

- $a_{ij} = 1$  si  $i = j$  ou  $j = n$ ,
  - $a_{ij} = -1$  si  $i > j$ ,
  - $a_{ij} = 0$  sinon.

Calculer la factorisation LU de  $\mathbb{A}$ .

## Correction

## Factorisation LU de la matrice A :

$$\left( \begin{array}{cccccc} 1 & 0 & \dots & \dots & 0 & 1 \\ -1 & 1 & \ddots & & \vdots & \vdots \\ \vdots & \ddots & 1 & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots & \vdots \\ \vdots & \ddots & \ddots & 1 & 1 & \\ -1 & \dots & \dots & \dots & -1 & 1 \end{array} \right) \xrightarrow{\substack{L_2 \leftarrow L_2 + L_1 \\ \vdots \\ L_n \leftarrow L_n + L_1}} \left( \begin{array}{cccccc} 1 & 0 & \dots & \dots & 0 & 1 \\ 0 & 1 & \ddots & & \vdots & 2 \\ \vdots & -1 & 1 & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 & \vdots \\ \vdots & \vdots & \ddots & 1 & 2 & \\ 0 & -1 & \dots & \dots & -1 & 2 \end{array} \right) \xrightarrow{\substack{L_3 \leftarrow L_3 + L_2 \\ \vdots \\ L_n \leftarrow L_n + L_2}} \left( \begin{array}{cccccc} 1 & 0 & \dots & \dots & 0 & 1 \\ 0 & 1 & \ddots & & \vdots & 2 \\ \vdots & 0 & 1 & \ddots & \vdots & 4 \\ \vdots & \vdots & -1 & \ddots & 0 & \vdots \\ \vdots & \vdots & \vdots & \ddots & 1 & 4 \\ 0 & 0 & -1 & \dots & -1 & 4 \end{array} \right)$$

[...]  $\xrightarrow{L_n \leftarrow L_n + L_{n-1}}$

$$\left( \begin{array}{cccccc} 1 & 0 & \dots & \dots & 0 & 2^0 \\ 0 & 1 & \ddots & & \vdots & 2^1 \\ \vdots & \ddots & 1 & \ddots & \vdots & 2^2 \\ \vdots & \ddots & \ddots & 0 & \vdots & \vdots \\ \vdots & & \ddots & 1 & 2^{n-2} & \\ 0 & \dots & \dots & \dots & 0 & 2^{n-1} \end{array} \right)$$

On obtient les matrices

$$\mathbb{L} = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & 0 \\ -1 & 1 & \ddots & & & \vdots \\ \vdots & \ddots & 1 & \ddots & & \vdots \\ & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & 1 & 0 \\ -1 & \dots & \dots & \dots & -1 & 1 \end{pmatrix} \quad \text{et} \quad \mathbb{U} = \begin{pmatrix} 1 & 0 & \dots & \dots & 0 & 2^0 \\ 0 & 1 & \ddots & & \vdots & 2^1 \\ \vdots & \ddots & 1 & \ddots & \vdots & 2^2 \\ & & \ddots & \ddots & 0 & \vdots \\ \vdots & & & \ddots & 1 & 2^{n-2} \\ 0 & \dots & \dots & \dots & 0 & 2^{n-1} \end{pmatrix}.$$

i.e.

- \*  $\ell_{ii} = 1$  pour  $i = 1, \dots, n$ ,
  - \*  $\ell_{ij} = -1$  si  $i > j$
  - \*  $\ell_{ij} = 0$  sinon;
  - \*  $u_{ii} = 1$  pour  $i = 1, \dots, n-1$ ,
  - \*  $u_{in} = 2^{i-1}$  pour  $i = 1, \dots, n$ ,
  - \*  $u_{ij} = 0$  sinon.

### Exercice 5.36

Calculer, lorsqu'il est possible, la factorisation LU des matrices suivantes :

$$\mathbb{A} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 7 & 8 & 9 \end{pmatrix}, \quad \mathbb{B} = \begin{pmatrix} 1 & 2 & 3 \\ 7 & 8 & 9 \\ 2 & 4 & 5 \end{pmatrix}.$$

Comment peut-on modifier l'algorithme de factorisation pour pouvoir toujours aboutir à une factorisation LU lorsque la matrice est inversible ?

#### Correction

Pour une matrice quelconque  $\mathbb{A} \in \mathcal{M}_{n,n}(\mathbb{R})$ , la factorisation LU (sans pivot) existe et est unique ssi les sous-matrices principales  $\mathbb{A}_i$  de  $\mathbb{A}$  d'ordre  $i = 1, \dots, n-1$  (celles que l'on obtient en restreignant  $\mathbb{A}$  à ses  $i$  premières lignes et colonnes) ne sont pas singulières (autrement dit si les mineurs principaux, i.e. les déterminants des sous-matrices principales, sont non nuls).

Matrice  $\mathbb{A}$  : comme  $\det(\mathbb{A}) \neq 0$ , la matrice  $\mathbb{A}$  est bien inversible. Puisque  $\det(\mathbb{A}_1) = a_{11} = 1 \neq 0$  mais  $\det(\mathbb{A}_2) = a_{11}a_{22} - a_{12}a_{21} = 0$ , on ne peut pas factoriser  $\mathbb{A}$  sans utiliser la technique du pivot. En effet,

$$\mathbb{A} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 7 & 8 & 9 \end{pmatrix} \xrightarrow[L_3 \leftarrow L_3 - \frac{7}{1}L_1]{L_2 \leftarrow L_2 - \frac{2}{1}L_1} \begin{pmatrix} 1 & 2 & 3 \\ 0 & 0 & -1 \\ 0 & -6 & -12 \end{pmatrix}$$

La factorisation LU ne peut pas être calculée car à la prochaine étape il faudrait effectuer le changement  $L_3 \leftarrow L_3 - \frac{-6}{0}L_2$ .

Matrice  $\mathbb{B}$  :

$$\mathbb{A}_2 = \begin{pmatrix} 1 & 2 & 3 \\ 7 & 8 & 9 \\ 2 & 4 & 5 \end{pmatrix} \xrightarrow[L_3 \leftarrow L_3 - \frac{2}{1}L_1]{L_2 \leftarrow L_2 - \frac{7}{1}L_1} \begin{pmatrix} 1 & 2 & 3 \\ 0 & -6 & -12 \\ 0 & 0 & -1 \end{pmatrix}$$

La factorisation LU de la matrice  $\mathbb{B}$  est donc

$$\mathbb{L} = \begin{pmatrix} 1 & 0 & 0 \\ 7 & 1 & 0 \\ 2 & 0 & 1 \end{pmatrix}, \quad \mathbb{U} = \begin{pmatrix} 1 & 2 & 3 \\ 0 & -6 & -12 \\ 0 & 0 & -1 \end{pmatrix}.$$

Lorsqu'un pivot est nul, la méthode de GAUSS pour calculer la factorisation LU de la matrice  $\mathbb{A}$  n'est plus applicable. De plus, si le pivot n'est pas nul mais très petit, l'algorithme conduit à des erreurs d'arrondi importantes. C'est pourquoi des algorithmes qui échangent les éléments de façon à avoir le pivot le plus grand possible ont été développés. Les programmes optimisés intervertissent les lignes à chaque étape de façon à placer en pivot le terme de coefficient le plus élevé : c'est la méthode du pivot partiel. Pour la matrice  $\mathbb{A}$  cela aurait donné

$$\mathbb{A} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 7 & 8 & 9 \end{pmatrix} \xrightarrow{L_2 \leftrightarrow L_3} \begin{pmatrix} 1 & 2 & 3 \\ 7 & 8 & 9 \\ 2 & 4 & 5 \end{pmatrix} \xrightarrow[L_3 \leftarrow L_3 - \frac{2}{1}L_1]{L_2 \leftarrow L_2 - \frac{7}{1}L_1} \begin{pmatrix} 1 & 2 & 3 \\ 0 & -6 & -12 \\ 0 & 0 & -1 \end{pmatrix}.$$

Bien évidemment, il faut garder trace de cet échange de lignes pour qu'il puisse être répercuté sur le terme source et sur l'inconnue lors de la résolution du système linéaire ; ceci est réalisé en introduisant une nouvelle matrice  $\mathbb{P}$ , dite matrice pivotale, telle que  $\mathbb{P}\mathbb{A} = \mathbb{LU}$  : la résolution du système linéaire  $\mathbb{A}\mathbf{x} = \mathbf{b}$  est donc ramené à la résolution des deux systèmes triangulaires  $\mathbb{L}\mathbf{y} = \mathbb{P}\mathbf{b}$  et  $\mathbb{U}\mathbf{x} = \mathbf{y}$ . Dans notre exemple cela donne

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

### Exercice 5.37

Soit  $\alpha$  un paramètre réel et soient les matrices  $\mathbb{A}_\alpha$ ,  $\mathbb{P}$  et le vecteur  $\mathbf{b}$  définis par

$$\mathbb{A}_\alpha = \begin{pmatrix} 2 & 4 & 1 \\ \alpha & -2 & -1 \\ 2 & 3 & 2 \end{pmatrix}, \quad \mathbb{P} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 0 \\ -3/2 \\ -1 \end{pmatrix}.$$

1. À quelle condition sur  $\alpha$ , la matrice  $\mathbb{A}_\alpha$  est inversible ?
2. À quelle condition sur  $\alpha$ , la matrice  $\mathbb{A}_\alpha$  admet-elle une décomposition LU (sans pivot) ?
3. Soit  $\alpha = -1$ . Calculer, si elle existe, la décomposition LU de la matrice  $\mathbb{M} = \mathbb{P}\mathbb{A}_\alpha$ .
4. Soit  $\alpha = -1$ . Résoudre le système linéaire  $\mathbb{A}\mathbf{x} = \mathbf{b}$  en résolvant le système linéaire  $\mathbb{M}\mathbf{x} = \mathbb{P}\mathbf{b}$ .

### Correction

1. La matrice  $\mathbb{A}_\alpha$  est inversible si et seulement si  $\det(\mathbb{A}) \neq 0$ . Comme

$$\begin{aligned}\det(\mathbb{A}) &= \det \begin{pmatrix} 2 & 4 & 1 \\ \alpha & -2 & -1 \\ 2 & 3 & 2 \end{pmatrix} \\ &= (2 \times (-2) \times 2) + (4 \times (-1) \times 2) + (1 \times \alpha \times 3) - (2 \times (-1) \times 3) - (4 \times \alpha \times 2) - (1 \times (-2) \times 2) \\ &= (-8) + (-8) + (3\alpha) - (-6) - (8\alpha) - (-4) \\ &= -6 - 5\alpha,\end{aligned}$$

la matrice  $\mathbb{A}_\alpha$  est inversible si et seulement si  $\alpha \neq -\frac{6}{5}$ .

2. Pour une matrice  $\mathbb{A}$  carrée d'ordre  $n$  quelconque, la factorisation de GAUSS existe et est unique si et seulement si les sous-matrices principales  $\mathbb{A}_i$  de  $\mathbb{A}$  d'ordre  $i = 1, \dots, n-1$  (celles que l'on obtient en restreignant  $\mathbb{A}$  à ses  $i$  premières lignes et colonnes) ne sont pas singulières (autrement dit si les mineurs principaux, *i.e.* les déterminants des sous-matrices principales, sont non nuls).

Pour la matrice  $\mathbb{A}_\alpha$  on a les sous-matrices principales suivantes :

$$\begin{aligned}\mathbb{A}_1 &= (2), & \det(\mathbb{A}_1) &= 2; \\ \mathbb{A}_2 &= \begin{pmatrix} 2 & 4 \\ \alpha & -2 \end{pmatrix}, & \det(\mathbb{A}_2) &= -4(1+\alpha).\end{aligned}$$

Par conséquent, la matrice  $\mathbb{A}_\alpha$  admet une décomposition LU (sans pivot) si et seulement si  $\alpha \neq -1$ .

3. Si  $\alpha = -1$  la matrice  $\mathbb{A}_\alpha$  n'admet pas de décomposition LU sans pivot. La matrice  $\mathbb{P}$  échange les lignes 2 et 3 de la matrice  $\mathbb{A}$  et on obtient la matrice

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

La matrice  $\mathbb{M}$  admet une décomposition LU (sans pivot) et l'on a

$$\begin{pmatrix} 2 & 4 & 1 \\ 2 & 3 & 2 \\ -1 & -2 & -1 \end{pmatrix} \xrightarrow[L_3 \leftarrow L_3 - \frac{-1}{2}L_1]{L_2 \leftarrow L_2 - L_1} \begin{pmatrix} 2 & 4 & 1 \\ 0 & -1 & 1 \\ 0 & 0 & -\frac{1}{2} \end{pmatrix}$$

Par conséquent, on obtient la décomposition LU suivante de la matrice  $\mathbb{M}$  :

$$\mathbb{L} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ -\frac{1}{2} & 0 & 1 \end{pmatrix}, \quad \mathbb{U} = \begin{pmatrix} 2 & 4 & 1 \\ 0 & -1 & 1 \\ 0 & 0 & -\frac{1}{2} \end{pmatrix}.$$

4. Pour résoudre le système linéaire  $\mathbb{M}\mathbf{x} = \mathbb{P}\mathbf{b}$  il suffit de résoudre les deux systèmes triangulaires suivants :

\*  $\mathbb{L}\mathbf{y} = \mathbb{P}\mathbf{b}$  :

$$y_1 = 0, \quad y_2 = -1 - y_1 = -1, \quad y_3 = -\frac{3}{2} + \frac{1}{2}y_1 = -\frac{3}{2};$$

\*  $\mathbb{U}\mathbf{x} = \mathbf{y}$  :

$$x_3 = \frac{-3}{2}(-2) = 3, \quad x_2 = (-1 - x_3)/(-1) = 4, \quad x_1 = (0 - x_2 - 4x_3)/2 = -\frac{19}{2}.$$

### Exercice 5.38

Considérons les deux matrices carrées d'ordre  $n > 3$  :

$$\mathbb{A} = \begin{pmatrix} \alpha & 0 & 0 & 0 & \dots & \beta \\ 0 & \alpha & 0 & 0 & 0 & \dots & \beta \\ 0 & 0 & \alpha & 0 & \ddots & & \vdots \\ 0 & & \ddots & \ddots & & \dots & \beta \\ \vdots & \vdots & & \ddots & 0 & \beta \\ 0 & 0 & & 0 & \alpha & \beta \\ \beta & \beta & \dots & \beta & \beta & \alpha \end{pmatrix} \quad \mathbb{B} = \begin{pmatrix} \beta & 0 & \dots & \dots & 0 & 0 & \alpha \\ \beta & 0 & 0 & 0 & 0 & \alpha & 0 \\ \vdots & & 0 & \ddots & & 0 \\ & & & \ddots & & \dots & \vdots \\ \vdots & 0 & \alpha & 0 & 0 & 0 & 0 \\ \beta & \alpha & 0 & 0 & 0 & \alpha & 0 \\ \alpha & \beta & \beta & \dots & \beta & \beta & \beta \end{pmatrix}$$

avec  $\alpha$  et  $\beta$  réels non nuls.

1. Vérifier que la factorisation LU de la matrice  $\mathbb{B}$  ne peut pas être calculée sans utiliser la technique du pivot.
2. Calculer analytiquement le nombre d'opérations nécessaires pour calculer la factorisation LU de la matrice  $\mathbb{A}$ .
3. Exprimer le déterminant de la matrice  $\mathbb{A}$  sous forme récursive en fonction des coefficients de la matrice et de sa dimension  $n$ .
4. Sous quelles conditions sur  $\alpha$  et  $\beta$  la matrice  $\mathbb{A}$  est définie positive ? Dans ce cas, exprimer le conditionnement de la matrice en fonction des coefficients et de la dimension  $n$ .

#### Correction

1. La factorisation LU de la matrice  $\mathbb{B}$  ne peut pas être calculée sans utiliser la technique du pivot car l'élément pivotale au deuxième pas est nul. Par exemple, si  $n = 4$ , on obtient :

$$\mathbb{B}^{(1)} = \begin{pmatrix} \beta & 0 & 0 & \alpha \\ \beta & 0 & \alpha & 0 \\ \beta & \alpha & 0 & 0 \\ \alpha & \beta & \beta & \beta \end{pmatrix} \xrightarrow[L_2 \leftarrow L_2 - L_1]{L_3 \leftarrow L_3 - L_1} \mathbb{B}^{(2)} = \begin{pmatrix} \beta & 0 & 0 & \alpha \\ 0 & \boxed{0} & \alpha & -\alpha \\ 0 & \alpha & 0 & -\alpha \\ 0 & \beta & \beta & \beta - \frac{\alpha^2}{\beta} \end{pmatrix}.$$

2. La matrice  $\mathbb{A}$  est une matrice «en flèche» : pour en calculer la factorisation LU il suffit de transformer la dernière ligne, ce qui requiert le calcul de l'unique multiplicateur  $\ell_{nk} = \beta/\alpha$  et l'exécution de  $n - 1$  produits et sommes. Le coût global est donc de l'ordre de  $n$ .
3. Le déterminant  $\delta_n$  de la matrice  $\mathbb{A}$  de dimension  $n$  coïncide avec le déterminant de la matrice U. Comme  $u_{ii} = \alpha$  pour tout  $i < n$  et  $u_{nn} = \alpha - (n-1)\beta^2/\alpha$ , on conclut que

$$\delta_n = \prod_{i=1}^n u_{ii} = u_{nn} \cdot \prod_{i=1}^{n-1} u_{ii} = \left( \alpha - (n-1) \frac{\beta^2}{\alpha} \right) \alpha^{n-1} = \alpha^n - (n-1)\alpha^{n-2}\beta^2.$$

4. Les valeurs propres de la matrice  $\mathbb{A}$  sont les racines du déterminant de la matrice  $\mathbb{A} - \lambda \mathbb{I}$ . Suivant le même raisonnement du point précédent, ce déterminant s'écrit

$$(\alpha - \lambda)^n - (n-1)(\alpha - \lambda)^{n-2}\beta^2$$

dont les racines sont

$$\lambda_{1,2} = \alpha \pm \sqrt{(n-1)\beta}, \quad \lambda_3 = \dots = \lambda_n = \alpha.$$

Par conséquent, pour que la matrice  $\mathbb{A}$  soit définie positive il faut que les valeurs propres soient tous positifs, ce qui impose

$$\alpha > 0, \quad |\beta| < \frac{\alpha}{\sqrt{n-1}}.$$

Dans ce cas, le conditionnement de la matrice en norme 2 est

$$K_2(\mathbb{A}) = \begin{cases} \frac{\alpha + \beta\sqrt{n-1}}{\alpha - \beta\sqrt{n-1}} & \text{si } \beta \geq 0, \\ \frac{\alpha - \beta\sqrt{n-1}}{\alpha + \beta\sqrt{n-1}} & \text{sinon.} \end{cases}$$

### Exercice 5.39

Écrire les formules de la méthode d'élimination de GAUSS pour une matrice de la forme

$$\mathbb{A} = \begin{pmatrix} a_{1,1} & a_{1,2} & 0 & \dots & 0 \\ a_{2,1} & a_{2,2} & a_{2,3} & 0 & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & 0 \\ a_{n,1} & a_{n,2} & \dots & a_{n,n-1} & a_{n-1,n} & a_{n,n} \end{pmatrix}.$$

Quelle est la forme finale de la matrice  $\mathbb{U} = \mathbb{A}^{(n)}$ ? Étant donné la forme particulière de la matrice  $\mathbb{A}$ , indiquer le nombre minimal d'opérations nécessaire pour calculer  $\mathbb{U}$  ainsi que celui pour la résolution des systèmes triangulaires finaux.

#### Correction

Comme la matrice a une seule sur-diagonale non nulle, les formules de la méthode d'élimination de GAUSS deviennent

$$\begin{aligned} a_{ij}^{(k+1)} &= a_{ij}^{(k)} + \ell_{ik} a_{kj}^{(k)}, & i, j = k+1, \\ \ell_{ik} &= \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}, & i = k+1. \end{aligned}$$

La coût est donc de l'ordre de  $n$  et la matrice  $\mathbb{U}$  est bidiagonale supérieure.

### Exercice 5.40

Soit  $\alpha \in \mathbb{R}^*$  et considérons les matrices carrées de dimension  $n$

$$\mathbb{A} = \begin{pmatrix} \alpha & 0 & \dots & -\alpha \\ 0 & \ddots & & \vdots \\ \vdots & & \alpha & -\alpha \\ -\alpha & \dots & -\alpha & -\alpha \end{pmatrix}, \quad \mathbb{B} = \begin{pmatrix} \frac{\beta}{\alpha} & -\frac{\gamma}{\alpha} & \dots & -\frac{\gamma}{\alpha} \\ -\frac{\gamma}{\alpha} & \ddots & & \vdots \\ \vdots & & \frac{\beta}{\alpha} & -\frac{\gamma}{\alpha} \\ -\frac{\gamma}{\alpha} & \dots & -\frac{\gamma}{\alpha} & \frac{\beta}{\alpha} \end{pmatrix}.$$

1. Calculer  $\gamma$  et  $\beta$  pour que  $\mathbb{B}$  soit l'inverse de  $\mathbb{A}$ .
2. Calculer le conditionnement  $K_\infty(\mathbb{A})$  en fonction de  $n$  et en calculer la limite pour  $n$  qui tend vers l'infini.

#### Correction

1. Par définition,  $\mathbb{B}$  est la matrice inverse de  $\mathbb{A}$  si  $\mathbb{A}\mathbb{B} = \mathbb{B}\mathbb{A} = \mathbb{I}$ . Comme

$$\mathbb{A}\mathbb{B} = \begin{pmatrix} \beta + \gamma & 0 & \dots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \beta + \gamma & 0 \\ -\beta + (n-3)\gamma & \dots & -\beta + (n-3)\gamma & (n-2)\gamma \end{pmatrix},$$

il faut que

$$\begin{cases} \beta + \gamma = 1 \\ -\beta + (n-3)\gamma = 0 \\ (n-2)\gamma = 1 \end{cases}$$

ce qui donne

$$\beta = \frac{n-3}{n-2}, \quad \gamma = \frac{1}{n-2}.$$

2. On trouve immédiatement  $\|\mathbb{A}\|_\infty = n|\alpha|$  tandis que

$$\|\mathbb{A}^{-1}\|_\infty = \frac{1}{|\alpha|} \max \left\{ n, \frac{n}{n-2} \right\} = \frac{2}{|\alpha|}.$$

On conclut que le conditionnement  $K_\infty(\mathbb{A})$  en fonction de  $n$  est

$$K_\infty(\mathbb{A}) = n|\alpha| \frac{2}{|\alpha|} = 2n.$$

La matrice est donc mal conditionnée pour  $n$  grand.

### Exercice 5.41

On suppose que le nombre réel  $\varepsilon > 0$  est assez petit pour que l'ordinateur arrondisse  $1 + \varepsilon$  en 1 et  $1 + (1/\varepsilon)$  en  $1/\varepsilon$  ( $\varepsilon$  est plus petit que l'erreur machine (relative), par exemple,  $\varepsilon = 2^{-30}$  en format 32 bits). Simuler la résolution par l'ordinateur des deux systèmes suivants :

$$\begin{cases} \varepsilon a + b = 1 \\ 2a + b = 0 \end{cases} \quad \text{et} \quad \begin{cases} 2a + b = 0 \\ \varepsilon a + b = 1 \end{cases}$$

On appliquera pour cela la méthode du pivot de GAUSS et on donnera les décompositions LU des deux matrices associées à ces systèmes. On fournira également la solution exacte de ces systèmes. Commenter.

#### Correction

Premier système :

$$\begin{pmatrix} \varepsilon & 1 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Factorisation LU :

$$\begin{pmatrix} \varepsilon & 1 \\ 2 & 1 \end{pmatrix} \xrightarrow{L_2 \leftarrow L_2 - \frac{2}{\varepsilon} L_1} \begin{pmatrix} \varepsilon & 1 \\ 0 & 1 - \frac{2}{\varepsilon} \end{pmatrix} \quad \text{donc} \quad L = \begin{pmatrix} 1 & 0 \\ \frac{2}{\varepsilon} & 1 \end{pmatrix}, \quad U = \begin{pmatrix} \varepsilon & 1 \\ 0 & 1 - \frac{2}{\varepsilon} \end{pmatrix}$$

Pour résoudre le système linéaire on résout les systèmes triangulaires  $L\mathbf{y} = \mathbf{b}$  et  $U\mathbf{x} = \mathbf{y}$  :

$$\begin{array}{lll} \begin{pmatrix} 1 & 0 \\ \frac{2}{\varepsilon} & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} & \Rightarrow & y_1 = 1, \quad y_2 = -\frac{2}{\varepsilon}; \\ \begin{pmatrix} \varepsilon & 1 \\ 0 & 1 - \frac{2}{\varepsilon} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 1 \\ -\frac{2}{\varepsilon} \end{pmatrix} & \Rightarrow & b = -\frac{2}{\varepsilon(1 - \frac{2}{\varepsilon})}, \quad a = \frac{1 + \frac{2}{\varepsilon(1 - \frac{2}{\varepsilon})}}{\varepsilon}. \end{array}$$

Mais avec l'ordinateur, comme  $1 + \varepsilon \approx 1$  et  $1 + (1/\varepsilon) \approx 1/\varepsilon$ , on obtient

$$\tilde{L} = \begin{pmatrix} 1 & 0 \\ \frac{2}{\varepsilon} & 1 \end{pmatrix} \quad \tilde{U} = \begin{pmatrix} \varepsilon & 1 \\ 0 & -\frac{2}{\varepsilon} \end{pmatrix}$$

Pour résoudre ce système linéaire approché on résout les systèmes triangulaires  $\tilde{L}\mathbf{y} = \mathbf{b}$  et  $\tilde{U}\mathbf{x} = \mathbf{y}$  :

$$\begin{array}{lll} \begin{pmatrix} 1 & 0 \\ \frac{2}{\varepsilon} & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} & \Rightarrow & y_1 = 1, \quad y_2 = -\frac{2}{\varepsilon}; \\ \begin{pmatrix} \varepsilon & 1 \\ 0 & -\frac{2}{\varepsilon} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 1 \\ -\frac{2}{\varepsilon} \end{pmatrix} & \Rightarrow & b = 1, \quad a = 0. \end{array}$$

Second système :

$$\begin{pmatrix} 2 & 1 \\ \varepsilon & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Factorisation LU :

$$\begin{pmatrix} 2 & 1 \\ \varepsilon & 1 \end{pmatrix} \xrightarrow{L_2 \leftarrow L_2 - \frac{\varepsilon}{2} L_1} \begin{pmatrix} 2 & 1 \\ 0 & 1 - \frac{\varepsilon}{2} \end{pmatrix} \quad \text{donc} \quad L = \begin{pmatrix} 1 & 0 \\ \frac{\varepsilon}{2} & 1 \end{pmatrix}, \quad U = \begin{pmatrix} 2 & 1 \\ 0 & 1 - \frac{\varepsilon}{2} \end{pmatrix}$$

Pour résoudre le système linéaire on résout les systèmes triangulaires  $L\mathbf{y} = \mathbf{b}$  et  $U\mathbf{x} = \mathbf{y}$  :

$$\begin{array}{lll} \begin{pmatrix} 1 & 0 \\ \frac{\varepsilon}{2} & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} & \Rightarrow & y_1 = 0, \quad y_2 = 1; \\ \begin{pmatrix} 2 & 1 \\ 0 & 1 - \frac{\varepsilon}{2} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} & \Rightarrow & b = -\frac{2}{\varepsilon(1 - \frac{\varepsilon}{2})}, \quad a = \frac{1 + \frac{2}{\varepsilon(1 - \frac{\varepsilon}{2})}}{\varepsilon}. \end{array}$$

Mais avec l'ordinateur, comme  $1 + \varepsilon \approx 1$  et  $1 + (1/\varepsilon) \approx 1/\varepsilon$ , on obtient

$$\tilde{\mathbb{L}} = \begin{pmatrix} 1 & 0 \\ \frac{\varepsilon}{2} & 1 \end{pmatrix} \quad \tilde{\mathbb{U}} = \begin{pmatrix} 2 & 1 \\ 0 & -\frac{2}{\varepsilon} \end{pmatrix}$$

Pour résoudre ce système linéaire approché on résout les systèmes triangulaires  $\tilde{\mathbb{L}}\mathbf{y} = \mathbf{b}$  et  $\tilde{\mathbb{U}}\mathbf{x} = \mathbf{y}$ :

$$\begin{aligned} \begin{pmatrix} 1 & 0 \\ \frac{\varepsilon}{2} & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} &= \begin{pmatrix} 0 \\ 1 \end{pmatrix} \implies y_1 = 0, \quad y_2 = 1; \\ \begin{pmatrix} 2 & 1 \\ 0 & -\frac{2}{\varepsilon} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} &= \begin{pmatrix} 0 \\ 1 \end{pmatrix} \implies b = -\frac{\varepsilon}{2}, \quad a = \frac{\varepsilon}{4}. \end{aligned}$$

### Exercice 5.42

Rappeler l'algorithme vu en cours pour calculer la décomposition  $\mathbb{LU}$  d'une matrice  $\mathbb{A}$  et la solution du système  $\mathbb{A}\mathbf{x} = \mathbf{b}$  où le vecteur colonne  $\mathbf{b}$  est donné. On appliquera ces algorithmes pour les cas suivants :

$$\begin{pmatrix} 1 & 1 & 1 \\ 2 & 1 & 3 \\ -3 & 2 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \quad \text{et} \quad \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & -5 & 7 & 1 \\ 3 & 1 & 1 & 5 \\ 2 & 2 & 0 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \quad \text{et} \quad \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -2 & 3 & 4 \\ 1 & 4 & 6 & 8 \\ 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

#### Correction

Premier système :

$$\left( \begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 2 & 1 & 3 & 1 \\ -3 & 2 & 4 & 1 \end{array} \right) \xrightarrow[L_2 \leftarrow L_2 - \frac{2}{1}L_1]{L_3 \leftarrow L_3 - \frac{-3}{1}L_1} \left( \begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & -1 & 1 & -1 \\ 0 & 5 & 7 & 4 \end{array} \right) \xrightarrow[L_3 \leftarrow L_3 - \frac{5}{-1}L_2]{L_1 \leftarrow L_1 - L_2} \left( \begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & -1 & 1 & -1 \\ 0 & 0 & 12 & -1 \end{array} \right)$$

donc

$$\mathbb{L} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -3 & -5 & 1 \end{pmatrix} \quad \mathbb{U} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & -1 & 1 \\ 0 & 0 & 12 \end{pmatrix}$$

Il ne reste à résoudre que le système triangulaire

$$\begin{cases} x_1 + x_2 + x_3 = 1 \\ -x_2 + x_3 = -1 \\ 12x_3 = -1 \end{cases} \implies x_3 = -\frac{1}{12}, \quad x_2 = \frac{11}{12}, \quad x_1 = \frac{1}{6}.$$

Deuxième système :

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 1 \\ 2 & -5 & 7 & 1 & 1 \\ 3 & 1 & 1 & 5 & 1 \\ 2 & 2 & 0 & 3 & 1 \end{pmatrix} \xrightarrow[L_2 \leftarrow L_2 - \frac{2}{1}L_1]{L_3 \leftarrow L_3 - \frac{3}{1}L_1} \left( \begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 0 & -9 & 1 & -7 & -1 \\ 0 & -5 & -8 & -7 & -2 \\ 0 & -2 & -6 & -5 & -1 \end{array} \right) \xrightarrow[L_3 \leftarrow L_3 - \frac{-5}{-9}L_2]{L_4 \leftarrow L_4 - \frac{-2}{-9}L_2} \left( \begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 0 & -9 & 1 & -7 & -1 \\ 0 & 0 & -\frac{77}{9} & -\frac{28}{9} & -\frac{13}{9} \\ 0 & 0 & -\frac{56}{9} & -\frac{31}{9} & -\frac{9}{9} \end{array} \right) \xrightarrow[L_4 \leftarrow L_4 - \frac{56/9}{77/9}L_2]{L_4 \leftarrow L_4 - \frac{56/9}{77/9}L_2} \left( \begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 0 & -9 & 1 & -7 & -1 \\ 0 & 0 & -\frac{77}{9} & -\frac{28}{9} & -\frac{13}{9} \\ 0 & 0 & 0 & -\frac{13}{11} & \frac{3}{11} \end{array} \right)$$

donc

$$\mathbb{L} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & \frac{5}{9} & 1 & 0 \\ 2 & \frac{2}{9} & \frac{56}{77} & 1 \end{pmatrix} \quad \mathbb{U} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & -9 & 1 & -7 \\ 0 & 0 & -\frac{77}{9} & -\frac{28}{9} \\ 0 & 0 & 0 & -\frac{13}{11} \end{pmatrix}$$

Il ne reste à résoudre que le système triangulaire

$$\begin{cases} x_1 + 2x_2 + 3x_3 + 4x_4 = 1 \\ -9x_2 + x_3 - 7x_4 = -1 \\ -\frac{77}{9}x_3 - \frac{28}{9}x_4 = -\frac{13}{9} \\ -\frac{13}{11}x_4 = \frac{3}{11} \end{cases} \implies x_4 = -\frac{3}{13}, \quad x_3 = \frac{23}{91}, \quad x_2 = \frac{29}{91}, \quad x_1 = \frac{48}{91}.$$

Troisième système :

$$\left( \begin{array}{cccc|c} 1 & 1 & 1 & 1 & 1 \\ 1 & -2 & 3 & 4 & 1 \\ 1 & 4 & 6 & 8 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{array} \right) \xrightarrow{\substack{L_2 \leftarrow L_2 - L_1 \\ L_3 \leftarrow L_3 - L_1 \\ L_4 \leftarrow L_4 - L_1}} \left( \begin{array}{cccc|c} 1 & 1 & 1 & 1 & 1 \\ 0 & -3 & 2 & 3 & -0 \\ 0 & 3 & 5 & 7 & 0 \\ 0 & -1 & -1 & -1 & 0 \end{array} \right) \xrightarrow{\substack{L_3 \leftarrow L_3 - (-1)L_2 \\ L_4 \leftarrow L_4 - \frac{-1}{-3}L_2}} \left( \begin{array}{cccc|c} 1 & 1 & 1 & 1 & 1 \\ 0 & -3 & 2 & 3 & -0 \\ 0 & 0 & 7 & 10 & 0 \\ 0 & 0 & -\frac{5}{3} & -2 & 0 \end{array} \right) \xrightarrow{L_4 \leftarrow L_4 - \frac{-5/3}{7}L_2} \left( \begin{array}{cccc|c} 1 & 1 & 1 & 1 & 1 \\ 0 & -3 & 2 & 3 & -0 \\ 0 & 0 & 7 & 10 & 0 \\ 0 & 0 & 0 & \frac{8}{21} & 0 \end{array} \right)$$

donc

$$\mathbb{L} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & -1 & 1 & 0 \\ 1 & \frac{1}{3} & \frac{-5}{21} & 1 \end{pmatrix} \quad \mathbb{U} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & -3 & 2 & 3 \\ 0 & 0 & 7 & 10 \\ 0 & 0 & 0 & \frac{8}{21} \end{pmatrix}$$

Il ne reste à résoudre que le système triangulaire

$$\begin{cases} x_1 + x_2 + x_3 + x_4 = 1 \\ -3x_2 + 2x_3 + 3x_4 = 0 \\ 7x_3 + 10x_4 = 0 \\ \frac{8}{21}x_4 = 0 \end{cases} \implies x_4 = 0, \quad x_3 = 0, \quad x_2 = 0, \quad x_1 = 1.$$

## Factorisation QR et systèmes linéaires sur déterminés

### ★ Exercice 5.43 (Système sur-déterminé)

Soit le système linéaire sur-déterminé  $\mathbb{A}\mathbf{x} = \mathbf{b}$  avec  $\mathbb{A}$  la matrice de 8 lignes et 2 colonnes et  $\mathbf{b}$  le vecteur de 8 lignes suivantes :

$$\mathbb{A} = \begin{pmatrix} 0 & 1 \\ 0.06 & 1 \\ 0.14 & 1 \\ 0.25 & 1 \\ 0.31 & 1 \\ 0.47 & 1 \\ 0.6 & 1 \\ 0.7 & 1 \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} 0 \\ 0.08 \\ 0.14 \\ 0.2 \\ 0.23 \\ 0.25 \\ 0.28 \\ 0.29 \end{pmatrix}$$

Calculer la solution  $\mathbf{x} \in \mathbb{R}^2$  au sens des moindres carrés en utilisant la factorisation QR. Comparer la solution obtenue en résolvant le système  $\mathbb{R}\mathbf{x} = \mathbb{Q}^T \mathbf{b}$  avec le système  $\tilde{\mathbb{R}}\mathbf{x} = \tilde{\mathbb{Q}}^T \mathbf{b}$  et avec la solution donnée par Octave  $\mathbb{A}\backslash\mathbf{b}$ .

TO DO : Faire le lien avec un pb de fitting affine

### Correction

```
A=[ 0 1; 0.06 1; 0.14 1; 0.25 1; 0.31 1; 0.47 1; 0.6 1; 0.7 1 ]
```

```
b=[0; 0.08; 0.14; 0.2; 0.23; 0.25; 0.28; 0.29]
```

```
[m,n]=size(A)
```

```
[Q,R]=qr(A)
```

```
xstar=R\ (Q'*b)
```

```
Qt=Q(:,1:n);
```

```
Rt=R(1:n,:);
```

```
xstar=Rt\(Qt'*b)
xstar=A\b
```

## Méthodes itératives

### ★ Exercice 5.44 (systèmes linéaires, méthodes itératives)

Une méthode itérative pour le calcul de la solution d'un système linéaire  $\mathbb{A}\mathbf{x} = \mathbf{b}$  avec  $\mathbb{A} \in \mathbb{R}^{n \times n}$  est une méthode qui construit une suite de vecteurs  $\mathbf{x}^{(k)} = (x_1^{(k)}, x_2^{(k)}, \dots, x_n^{(k)})^T \in \mathbb{R}^n$  convergent vers le vecteur solution exacte  $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$  pour tout vecteur initiale  $\mathbf{x}^{(0)} = (x_1^{(0)}, x_2^{(0)}, \dots, x_n^{(0)})^T \in \mathbb{R}^n$  lorsque  $k$  tend vers  $+\infty$ .

**Méthode de Jacobi** Soit  $\mathbf{x}^0 = (x_1^0, x_2^0, \dots, x_n^0)$  un vecteur donné. La méthode de JACOBI définit la composante  $x_i^{k+1}$  du vecteur  $\mathbf{x}^{k+1}$  à partir des composantes  $x_j^k$  du vecteur  $\mathbf{x}^k$  pour  $j \neq i$  de la manière suivante :

$$x_i^{k+1} = \frac{b_i - \sum_{\substack{j=1 \\ j \neq i}}^n a_{ij} x_j^k}{a_{ii}}, \quad i = 1, \dots, n$$

Si la matrice  $\mathbb{A}$  est à diagonale dominante stricte, la méthode de JACOBI converge.

**Méthode de Gauss-Sidel** C'est une amélioration de la méthode de JACOBI dans laquelle les valeurs calculées sont utilisées au fur et à mesure du calcul et non à l'issue d'une itération comme dans la méthode de JACOBI. Soit  $\mathbf{x}^0 = (x_1^0, x_2^0, \dots, x_n^0)$  un vecteur donné. La méthode de GAUSS-SIDEL définit la composante  $x_i^{k+1}$  du vecteur  $\mathbf{x}^{k+1}$  à partir des composantes  $x_j^{k+1}$  du vecteur  $\mathbf{x}^{k+1}$  pour  $j < i$  et des composantes  $x_j^k$  du vecteur  $\mathbf{x}^k$  pour  $j \geq i$  de la manière suivante :

$$x_i^{k+1} = \frac{b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{k+1} - \sum_{j=i+1}^n a_{ij} x_j^k}{a_{ii}}, \quad i = 1, \dots, n$$

1. Implémenter une fonction appelée `myJacobi` permettant de résoudre un système linéaire  $\mathbb{A}\mathbf{x} = \mathbf{b}$  d'inconnue  $\mathbf{x}$  par la méthode itérative de Jacobi. La syntaxe doit être `[x, r, k]=myJacobi(A, b, xinit, toll, kmax)` où  $\mathbb{A}$  est une matrice de  $\mathbb{R}^{n \times n}$ ,  $\mathbf{b}$  est un vecteur colonne de  $\mathbb{R}^n$ ,  $xinit = \mathbf{x}^{(0)}$  est un vecteur colonne de  $\mathbb{R}^n$ ,  $toll$  la tolérance sur la norme du résidu  $\mathbb{A}\mathbf{x} - \mathbf{b}$  et  $kmax$  le nombre maximal d'itérations. On doit obtenir  $\mathbf{x}$  un vecteur colonne de  $\mathbb{R}^n$  solution du système linéaire  $\mathbb{A}\mathbf{x} = \mathbf{b}$ ,  $r$  la norme du dernier résidu calculé et  $k$  le nombre d'itérations effectuées.

Écrire un script appelé `TESTmyJacobi.m` pour tester cette fonction sur l'exemple suivant :  $toll = 10^{-9}$ ,  $kmax = 50$ ,

$$\mathbb{A} = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \mathbf{x}^{(0)} = \begin{pmatrix} 1 \\ 1/2 \end{pmatrix}.$$

La solution exacte est

$$\mathbf{x} = \begin{pmatrix} 3/5 \\ -1/5 \end{pmatrix}.$$

Construire la matrice d'itération associée à la méthode de Jacobi et en calculer le rayon spectrale.

2. Même exercice pour la méthode de Gauss-Seidel.

### Correction

1. Dans le fichier `myJacobi.m` on écrit

```
function [x,r,k]=myJacobi(A,b,xinit,toll,kmax)
k=0;
xold=xinit;
r=norm(A*xold'-b);
n=length(b);
while r>=toll & k<=kmax
    for i=1:n
        xnew(i)=(b(i)-dot(A(i,[1:i-1,i+1:n]),xold([1:i-1,i+1:n])))/A(i,i);
    end
    k+=1;
    if k>kmax
        break;
    end
    if abs(r)<toll
        break;
    end
    xold=xnew;
    r=norm(A*xold'-b);
end
```

```

xold=xnew;
r=norm(A*xold'-b);
end
x=xnew;
end

```

et on teste cette fonction par exemple comme suit

```

clear all
A=[2 1; 1 3]
b=[1;0]
xinit=[0 0];
[x,r,k]=myJacobi(A,b,xinit,1.e-9,50)

E=-tril(A,-1);
F=-triu(A,1);
P=A+E+F;
B=inv(P)*(P-A);
RayonSpectraleB_Jacobi=max(abs(eig(B)))

```

2. Dans le fichier myGS.m on écrit

```

function [x,r,k]=myGS(A,b,xinit,toll,kmax)
k=0;
x=xinit;
r=norm(A*x'-b);
n=length(b);
while r>=toll & k<=kmax
    for i=1:n
        x(i)=(b(i)-dot(A(i,[1:i-1,i+1:n]),x([1:i-1,i+1:n])))/A(i,i);
    end
    k+=1;
    r=norm(A*x'-b);
end
end

```

et on teste cette fonction par exemple comme suit

```

A=[2 1; 1 3]
b=[1;0]
xinit=[0 0];
[x,r,k]=myGS(A,b,xinit,1.e-9,50)

E=-tril(A,-1);
F=-triu(A,1);
P=A+E+F;
B=inv(P-E)*(P-E-A);
RayonSpectraleB_GS=max(abs(eig(B)))

```

### Exercice 5.45

Soit le système linéaire

$$\begin{pmatrix} 6 & 1 & 1 \\ 2 & 4 & 0 \\ 1 & 2 & 6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 12 \\ 0 \\ 6 \end{pmatrix}.$$

1. Approcher la solution avec la méthode de JACOBI avec 3 itérations à partir de  $\mathbf{x}^{(0)} = (2, 2, 2)$ .
2. Approcher la solution avec la méthode de GAUSS-SEIDEL avec 3 itérations à partir de  $\mathbf{x}^{(0)} = (2, 2, 2)$ .
3. Résoudre les systèmes linéaires par la méthode d'élimination de GAUSS.
4. Factoriser la matrice  $A$  (sans utiliser la technique du pivot) et résoudre les systèmes linéaires.

### Correction

1. Méthode de JACOBI :

$$\mathbf{x}^{(0)} = \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix}, \quad \mathbf{x}^{(1)} = \begin{pmatrix} \frac{12-(1\times 2+1\times 2)}{6} \\ \frac{0-(2\times 2+0\times 2)}{4} \\ \frac{6-(1\times 2+2\times 2)}{6} \end{pmatrix} = \begin{pmatrix} 4/3 \\ -1 \\ 0 \end{pmatrix}, \quad \mathbf{x}^{(2)} = \begin{pmatrix} \frac{12-(1\times(-1)+1\times 0)}{6} \\ \frac{0-(2\times\frac{4}{3}+0\times 0)}{4} \\ \frac{6-(1\times\frac{4}{3}+2\times(-1))}{6} \end{pmatrix} = \begin{pmatrix} 13/6 \\ -2/3 \\ 10/9 \end{pmatrix}, \quad \mathbf{x}^{(3)} = \begin{pmatrix} \frac{12-(1\times\frac{-2}{3}+1\times\frac{10}{9})}{6} \\ \frac{0-(2\times\frac{13}{6}+0\times\frac{10}{9})}{4} \\ \frac{6-(1\times\frac{13}{6}+2\times\frac{-2}{3})}{6} \end{pmatrix} = \begin{pmatrix} 52/27 \\ -13/12 \\ 31/36 \end{pmatrix}$$

ainsi

$$\mathbf{x} \approx \begin{pmatrix} 1.926 \\ -1.083 \\ 0.861 \end{pmatrix}.$$

2. Méthode de GAUSS-SEIDEL :

$$\mathbf{x}^{(0)} = \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix}, \quad \mathbf{x}^{(1)} = \begin{pmatrix} \frac{12-(1 \times 2+1 \times 2)}{6} \\ \frac{0-(2 \times \frac{4}{3}+0 \times 2)}{\frac{4}{3}} \\ \frac{6-(1 \times \frac{4}{3}+2 \times \frac{-2}{3})}{6} \end{pmatrix} = \begin{pmatrix} 4/3 \\ -\frac{2}{3} \\ 1 \end{pmatrix}, \quad \mathbf{x}^{(2)} = \begin{pmatrix} \frac{12-(1 \times \frac{-2}{3}+1 \times 1)}{6} \\ \frac{0-(2 \times \frac{35}{18}+0 \times 1)}{\frac{35}{18}} \\ \frac{6-(1 \times \frac{35}{18}+2 \times \frac{-35}{36})}{6} \end{pmatrix} = \begin{pmatrix} \frac{35}{18} \\ -\frac{35}{36} \\ 1 \end{pmatrix}, \quad \mathbf{x}^{(3)} = \begin{pmatrix} \frac{12-(1 \times \frac{35}{18}+1 \times \frac{-35}{36})}{6} \\ \frac{0-(2 \times \frac{431}{216}+0 \times 1)}{\frac{431}{216}} \\ \frac{6-(1 \times \frac{431}{216}+2 \times \frac{-431}{432})}{6} \end{pmatrix} = \begin{pmatrix} \frac{431}{216} \\ -\frac{431}{432} \\ 1 \end{pmatrix}$$

ainsi

$$\mathbf{x} \approx \begin{pmatrix} 1.995 \\ -0.995 \\ 1 \end{pmatrix}.$$

3. Méthode d'élimination de GAUSS :

$$(\mathbb{A}|\mathbf{b}) = \left( \begin{array}{ccc|c} 6 & 1 & 1 & 12 \\ 2 & 4 & 0 & 0 \\ 1 & 2 & 6 & 6 \end{array} \right) \xrightarrow{L_2 \leftarrow L_2 - \frac{2}{6}L_1} \left( \begin{array}{ccc|c} 6 & 1 & 1 & 12 \\ 0 & \frac{11}{3} & -\frac{1}{3} & -4 \\ 0 & \frac{11}{6} & \frac{35}{6} & 4 \end{array} \right) \xrightarrow{L_3 \leftarrow L_3 - \frac{11}{3}L_2} \left( \begin{array}{ccc|c} 6 & 1 & 1 & 12 \\ 0 & \frac{11}{3} & -\frac{1}{3} & -4 \\ 0 & 0 & 6 & 6 \end{array} \right)$$

donc

$$\begin{cases} 6x_1 + x_2 + x_3 = 12, \\ \frac{11}{3}x_2 - \frac{1}{3}x_3 = -4 \\ 6x_3 = 6 \end{cases} \implies x_3 = 1, \quad x_2 = -1, \quad x_1 = 2.$$

4. Factorisation de la matrice  $\mathbb{A}$  :

$$\left( \begin{array}{ccc} 6 & 1 & 1 \\ 2 & 4 & 0 \\ 1 & 2 & 6 \end{array} \right) \xrightarrow{L_2 \leftarrow L_2 - \frac{2}{6}L_1} \left( \begin{array}{ccc} 6 & 1 & 1 \\ 0 & \frac{11}{3} & -\frac{1}{3} \\ 0 & \frac{11}{6} & \frac{35}{6} \end{array} \right) \xrightarrow{L_3 \leftarrow L_3 - \frac{11}{3}L_2} \left( \begin{array}{ccc} 6 & 1 & 1 \\ 0 & \frac{11}{3} & -\frac{1}{3} \\ 0 & 0 & 6 \end{array} \right)$$

donc

$$\mathbb{L} = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{3} & 1 & 0 \\ \frac{1}{6} & \frac{1}{2} & 1 \end{pmatrix} \quad \mathbb{U} = \begin{pmatrix} 6 & 1 & 1 \\ 0 & \frac{11}{3} & -\frac{1}{3} \\ 0 & 0 & 6 \end{pmatrix}$$

Pour résoudre le système linéaire on résout les systèmes triangulaires  $\mathbb{L}\mathbf{y} = \mathbf{b}$

$$\begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{3} & 1 & 0 \\ \frac{1}{6} & \frac{1}{2} & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 12 \\ 0 \\ 6 \end{pmatrix} \implies y_1 = 12, \quad y_2 = -4, \quad y_3 = 6$$

et  $\mathbb{U}\mathbf{x} = \mathbf{y}$

$$\begin{pmatrix} 6 & 1 & 1 \\ 0 & \frac{11}{3} & -\frac{1}{3} \\ 0 & 0 & 6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ -4 \\ 6 \end{pmatrix} \implies x_3 = 1, \quad x_2 = -1, \quad x_1 = 2.$$

### Exercice 5.46

Donner une condition suffisante sur le coefficient  $\alpha$  pour avoir convergence des méthodes de JACOBI et GAUSS-SEIDEL pour la résolution d'un système linéaire associé à la matrice

$$\mathbb{A} = \begin{pmatrix} \alpha & 0 & 1 \\ 0 & \alpha & 0 \\ 1 & 0 & \alpha \end{pmatrix}$$

### Correction

Une condition suffisante pour la convergence des méthodes de JACOBI et de GAUSS-SEIDEL est que  $\mathbb{A}$  est à diagonale

strictement dominante, i.e.  $\sum_{\substack{i=1 \\ i \neq j}}^3 |a_{ij}| < |a_{ii}|$  pour  $j = 1, 2, 3$ . La matrice  $\mathbb{A}$  vérifie cette condition si et seulement si  $|\alpha| > 1$ .

### Exercice 5.47

Considérons le système linéaire  $\mathbb{A}\mathbf{x} = \mathbf{b}$  avec

$$\mathbb{A} = \begin{pmatrix} \alpha & 0 & \gamma \\ 0 & \alpha & \beta \\ 0 & \delta & \alpha \end{pmatrix}$$

avec  $\alpha, \beta, \gamma$  et  $\delta$  des paramètres réels. Donner des conditions suffisantes sur les coefficients pour avoir

1. convergence de la méthode de JACOBI
2. convergence de la méthode de GAUSS-SEIDEL.

#### Correction

1. Une condition suffisante pour que la méthode de JACOBI converge est que la matrice soit à dominance diagonale stricte, ce qui équivaut à imposer

$$\begin{cases} |\alpha| > |\gamma|, \\ |\alpha| > |\beta|, \\ |\alpha| > |\delta|, \end{cases}$$

c'est-à-dire  $|\alpha| > \max\{|\beta|, |\gamma|, |\delta|\}$ .

2. La condition précédente est aussi suffisante pour la convergence de la méthode de GAUSS-SEIDEL. Une autre condition suffisante pour la convergence de cette méthode est que la matrice soit symétrique définie positive. Pour la symétrie il faut que

$$\begin{cases} \gamma = 0, \\ \beta = \delta, \end{cases}$$

on obtient ainsi la matrice

$$\mathbb{A} = \begin{pmatrix} \alpha & 0 & 0 \\ 0 & \alpha & \beta \\ 0 & \beta & \alpha \end{pmatrix}.$$

Elle est définie positive si ses valeurs propres sont positifs. On a

$$\lambda_1 = \alpha, \quad \lambda_2 = \alpha - \beta, \quad \lambda_3 = \alpha + \beta,$$

donc il faut que  $\alpha > |\beta|$ .

On note que dans ce cas, lorsque  $\mathbb{A}$  est symétrique définie positive alors elle est aussi à dominance diagonale stricte.

### Exercice 5.48

Écrire les méthodes itératives de GAUSS, JACOBI et GAUSS-SEIDEL pour les systèmes suivants :

$$\begin{cases} 10a + b = 11 \\ 2a + 10b = 12 \end{cases} \quad \text{et} \quad \begin{cases} 2a + 10b = 12 \\ 10a + b = 11. \end{cases}$$

Pour chacun de ces méthodes et systèmes, on illustrera les résultats théoriques de convergence/non-convergence en calculant les 3 premières itérés en prenant comme point de départ le vecteur  $(a, b) = (0, 0)$ .

#### Correction

**Gauss** \* Premier système :

$$\left( \begin{array}{cc|c} 10 & 1 & 11 \\ 2 & 10 & 12 \end{array} \right) \xrightarrow{L_2 \leftarrow L_2 - \frac{1}{10}L_1} \left( \begin{array}{cc|c} 10 & 1 & 11 \\ 0 & \frac{49}{5} & \frac{49}{5} \end{array} \right) \Rightarrow \begin{cases} 10a + b = 11 \\ \frac{49}{5}b = \frac{49}{5} \end{cases} \Rightarrow \begin{cases} a = 1 \\ b = 1. \end{cases}$$

\* Second système :

$$\left( \begin{array}{cc|c} 2 & 10 & 12 \\ 10 & 1 & 11 \end{array} \right) \xrightarrow{L_2 \leftarrow L_2 - \frac{10}{2}L_1} \left( \begin{array}{cc|c} 2 & 10 & 12 \\ 0 & -49 & -49 \end{array} \right) \Rightarrow \begin{cases} 2a + 10b = 12 \\ -49b = -49 \end{cases} \Rightarrow \begin{cases} a = 1 \\ b = 1. \end{cases}$$

**Jacobi**    \* Premier système :

$$\begin{cases} 10a + b = 11 \\ 2a + 10b = 12 \end{cases} \iff \begin{cases} a = \frac{11-b}{10} \\ b = \frac{12-2a}{10} \end{cases}$$

La matrice étant à diagonale dominante stricte, la méthode converge et on a

$$\mathbf{x}^{(0)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \mathbf{x}^{(1)} = \begin{pmatrix} \frac{11-0}{10} \\ \frac{12-0}{10} \end{pmatrix} = \begin{pmatrix} 11/10 \\ 12/10 \end{pmatrix}, \quad \mathbf{x}^{(2)} = \begin{pmatrix} \frac{11-\frac{12}{10}}{10} \\ \frac{12-2\frac{11}{10}}{10} \end{pmatrix} = \begin{pmatrix} 49/50 \\ 49/50 \end{pmatrix}, \quad \mathbf{x}^{(3)} = \begin{pmatrix} \frac{11-\frac{49}{50}}{10} \\ \frac{12-2\frac{49}{50}}{10} \end{pmatrix} = \begin{pmatrix} 501/500 \\ 502/500 \end{pmatrix}.$$

\* Second système :

$$\begin{cases} 2a + 10b = 12 \\ 10a + b = 11 \end{cases} \iff \begin{cases} a = \frac{12-10b}{2} \\ b = 11 - 10a \end{cases}$$

La méthode ne converge pas, en effet on a

$$\mathbf{x}^{(0)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \mathbf{x}^{(1)} = \begin{pmatrix} \frac{12-0}{2} \\ 11-0 \end{pmatrix} = \begin{pmatrix} 6 \\ 11 \end{pmatrix}, \quad \mathbf{x}^{(2)} = \begin{pmatrix} \frac{12-10 \times 11}{2} \\ 11 - 10 \times 6 \end{pmatrix} = \begin{pmatrix} -49 \\ -49 \end{pmatrix}, \quad \mathbf{x}^{(3)} = \begin{pmatrix} \frac{12-10 \times (-49)}{2} \\ 11 - 10 \times (-49) \end{pmatrix} = \begin{pmatrix} 251 \\ 501 \end{pmatrix}.$$

**Gauss-Seidel**    \* Premier système :

$$\begin{cases} 10a + b = 11 \\ 2a + 10b = 12 \end{cases} \iff \begin{cases} a = \frac{11-b}{10} \\ b = \frac{12-2a}{10} \end{cases}$$

La matrice étant à diagonale dominante stricte, la méthode converge et on a

$$\mathbf{x}^{(0)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \mathbf{x}^{(1)} = \begin{pmatrix} \frac{11-0}{10} \\ \frac{12-2\frac{11}{10}}{10} \end{pmatrix} = \begin{pmatrix} 11/10 \\ 49/50 \end{pmatrix}, \quad \mathbf{x}^{(2)} = \begin{pmatrix} \frac{11-\frac{49}{50}}{10} \\ \frac{12-2\frac{49}{50}}{10} \end{pmatrix} = \begin{pmatrix} 501/500 \\ 2499/2500 \end{pmatrix}, \quad \mathbf{x}^{(3)} = \begin{pmatrix} \frac{11-\frac{2499}{2500}}{10} \\ \frac{12-2\frac{2499}{2500}}{10} \end{pmatrix} = \begin{pmatrix} 25001/25000 \\ 12499/125000 \end{pmatrix}.$$

\* Second système :

$$\begin{cases} 2a + 10b = 12 \\ 10a + b = 11 \end{cases} \iff \begin{cases} a = \frac{12-10b}{2} \\ b = 11 - 10a \end{cases}$$

La méthode ne converge pas, en effet on a

$$\mathbf{x}^{(0)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \mathbf{x}^{(1)} = \begin{pmatrix} \frac{12-0}{2} \\ 11 - 10 \times 6 \end{pmatrix} = \begin{pmatrix} 6 \\ -49 \end{pmatrix}, \quad \mathbf{x}^{(2)} = \begin{pmatrix} \frac{12-10 \times (-49)}{2} \\ 11 - 10 \times 251 \end{pmatrix} = \begin{pmatrix} 251 \\ -2499 \end{pmatrix}, \quad \mathbf{x}^{(3)} = \begin{pmatrix} \frac{12-10 \times (-2499)}{2} \\ 11 - 10 \times (12501) \end{pmatrix} = \begin{pmatrix} 12501 \\ -124999 \end{pmatrix}.$$

### Exercice 5.49

Soit  $\mathbb{A}$  une matrice,  $\mathbb{A} \in \mathcal{M}_{n,n}(\mathbb{R})$ .

1. Rappeler la méthode de JACOBI pour la résolution du système  $\mathbb{A}\mathbf{x} = \mathbf{b}$ , avec  $\mathbf{b} \in \mathcal{M}_{n,1}(\mathbb{R})$  donné.

2. Soit la matrice  $\mathbb{A}$  suivante :

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

La méthode de JACOBI est-elle convergente pour cette matrice ?

3. Construire à la main les matrices  $\mathbb{L}$  et  $\mathbb{U}$  de la factorisation  $\mathbb{L}\mathbb{U}$  pour la matrice ci-dessus.

### Correction

1. La méthode de JACOBI est une méthode itérative pour le calcul de la solution d'un système linéaire qui construit une suite de vecteurs  $\mathbf{x}^{(k)} \in \mathbb{R}^n$  convergent vers la solution exacte  $\mathbf{x}$  pour tout vecteur initiale  $\mathbf{x}^{(0)} \in \mathbb{R}^n$  :

$$x_i^{k+1} = \frac{b_i - \sum_{j=1}^n a_{ij} x_j^k}{a_{ii}}, \quad i = 1, \dots, n.$$

2. Comme  $|4| > |-1| + |-1|$ ,  $|3| > |-1| + |-1|$  et  $|4| > |-1| + |-1|$ , la matrice  $\mathbb{A}$  est à diagonale dominante stricte donc la méthode de JACOBI converge

3. Factorisation :

$$\left( \begin{array}{ccc} 4 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 4 \end{array} \right) \xrightarrow{\substack{L_2 \leftarrow L_2 - \frac{-1}{4}L_1 \\ L_3 \leftarrow L_3 - \frac{-1}{4}L_1}} \left( \begin{array}{ccc} 4 & -1 & -1 \\ 0 & 11/4 & -5/4 \\ 0 & -5/4 & 15/4 \end{array} \right) \xrightarrow{L_3 \leftarrow L_3 - \frac{-5/4}{11/4}L_2} \left( \begin{array}{ccc} 4 & -1 & -1 \\ 0 & 11/4 & -5/4 \\ 0 & 0 & 35/11 \end{array} \right).$$

Par conséquent

$$\mathbb{L} = \begin{pmatrix} 1 & 0 & 0 \\ -1/4 & 1 & 0 \\ -1/4 & -5/11 & 1 \end{pmatrix} \quad \text{et} \quad \mathbb{U} = \begin{pmatrix} 4 & -1 & -1 \\ 0 & 11/4 & -5/4 \\ 0 & 0 & 35/11 \end{pmatrix}.$$

### Exercice 5.50

Soit les systèmes linéaires

$$\begin{cases} 4x_1 + 3x_2 + 3x_3 = 10 \\ 3x_1 + 4x_2 + 3x_3 = 10 \\ 3x_1 + 3x_2 + 4x_3 = 10 \end{cases} \quad (5.1)$$

$$\begin{cases} 4x_1 + x_2 + x_3 = 6 \\ x_1 + 4x_2 + x_3 = 6 \\ x_1 + x_2 + 4x_3 = 6 \end{cases} \quad (5.2)$$

1. Rappeler une condition suffisante de convergence pour les méthodes de JACOBI et de GAUSS-SEIDEL. Rappeler une autre condition suffisante de convergence pour la méthode de GAUSS-SEIDEL (mais non pour la méthode de JACOBI). Les systèmes (5.1) et (5.2) vérifient-ils ces conditions ?
2. Écrire les méthodes de JACOBI et de GAUSS-SEIDEL pour ces deux systèmes linéaires.
3. On illustrera les résultats théoriques de convergence/non-convergence de ces deux schémas en prenant comme point de départ le vecteur  $(x_1, x_2, x_3) = (0, 0, 0)$  et en calculant les 3 premiers itérés :
  - 3.1. avec la méthode de JACOBI pour le système (5.1),
  - 3.2. avec la méthode de GAUSS-SEIDEL pour le système (5.1),
  - 3.3. avec la méthode de JACOBI pour le système (5.2),
  - 3.4. avec la méthode de GAUSS-SEIDEL pour le système (5.2).
4. On comparera le résultat obtenu avec la solution exacte (qu'on calculera à l'aide de la méthode d'élimination de GAUSS).

### Correction

Écrivons les deux systèmes sous forme matricielle  $\mathbb{A}\mathbf{x} = \mathbf{b}$  :

$$\underbrace{\begin{pmatrix} 4 & 3 & 3 \\ 3 & 4 & 3 \\ 3 & 3 & 4 \end{pmatrix}}_{\mathbb{A}_1} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 10 \\ 10 \\ 10 \end{pmatrix} \quad \text{et} \quad \underbrace{\begin{pmatrix} 4 & 1 & 1 \\ 1 & 4 & 1 \\ 1 & 1 & 4 \end{pmatrix}}_{\mathbb{A}_2} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 6 \\ 6 \\ 6 \end{pmatrix}$$

1. Rappelons deux propriétés de convergence :

- ★ Si la matrice  $\mathbb{A}$  est à diagonale dominante stricte, les méthodes de JACOBI et de GAUSS-SEIDEL convergent.
- ★ Si la matrice  $\mathbb{A}$  est symétrique et définie positive, la méthode de GAUSS-SEIDEL converge.

Comme  $4 > 1 + 1$ , la matrice  $\mathbb{A}_2$  est à diagonale dominante stricte : les méthodes de JACOBI et de GAUSS-SEIDEL convergent.

Comme  $4 < 3 + 3$ , la matrice  $\mathbb{A}_1$  n'est pas à diagonale dominante stricte : les méthodes de JACOBI et de GAUSS-SEIDEL peuvent ne pas converger. Cependant elle est symétrique et définie positive (car les valeurs propres<sup>4</sup> sont  $\lambda_1 = \lambda_2 = 1$  et  $\lambda_3 = 10$ ) : la méthode de GAUSS-SEIDEL converge.

2. Pour les systèmes donnés les méthodes de JACOBI et GAUSS-SEIDEL s'écrivent

---

4.  $\det \mathbb{A}_1(\lambda) = (4 - \lambda)^3 + 27 + 27 - 9(4 - \lambda) - 9(4 - \lambda) - 9(4 - \lambda) = 64 - 48\lambda + 12\lambda^2 - \lambda^3 + 54 - 108 + 27\lambda = -\lambda^3 + 12\lambda^2 - 21\lambda + 10$ . Une racine évidente est  $\lambda = 1$  et on obtient  $\det \mathbb{A}_1(\lambda) = (\lambda - 1)(-\lambda^2 + 11\lambda - 10) = (\lambda - 1)^2(\lambda - 10)$ .

	$\mathbb{A}_1 \mathbf{x} = \mathbf{b}$	$\mathbb{A}_2 \mathbf{x} = \mathbf{b}$
JACOBI	$\begin{pmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \\ x_3^{(k+1)} \end{pmatrix} = \frac{1}{4} \begin{pmatrix} 10 - 3x_2^{(k)} - 3x_3^{(k)} \\ 10 - 3x_1^{(k)} - 3x_3^{(k)} \\ 10 - 3x_1^{(k)} - 3x_2^{(k)} \end{pmatrix}$	$\begin{pmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \\ x_3^{(k+1)} \end{pmatrix} = \frac{1}{4} \begin{pmatrix} 6 - x_2^{(k)} - x_3^{(k)} \\ 6 - x_1^{(k)} - x_3^{(k)} \\ 6 - x_1^{(k)} - x_2^{(k)} \end{pmatrix}$
Gauss-SEIDEL	$\begin{pmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \\ x_3^{(k+1)} \end{pmatrix} = \frac{1}{4} \begin{pmatrix} 10 - 3x_2^{(k)} - 3x_3^{(k)} \\ 10 - 3x_1^{(k+1)} - 3x_3^{(k)} \\ 10 - 3x_1^{(k+1)} - 3x_2^{(k+1)} \end{pmatrix}$	$\begin{pmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \\ x_3^{(k+1)} \end{pmatrix} = \frac{1}{4} \begin{pmatrix} 6 - x_2^{(k)} - x_3^{(k)} \\ 6 - x_1^{(k+1)} - x_3^{(k)} \\ 6 - x_1^{(k+1)} - x_2^{(k+1)} \end{pmatrix}$

3. On obtient les suites suivantes

3.1. JACOBI pour le système (5.1) :

$$\begin{aligned} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(0)} &= \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \Rightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(1)} = \frac{1}{4} \begin{pmatrix} 10 - 3 \times 0 - 3 \times 0 \\ 10 - 3 \times 0 - 3 \times 0 \\ 10 - 3 \times 0 - 3 \times 0 \end{pmatrix} = \begin{pmatrix} \frac{5}{2} \\ \frac{5}{2} \\ \frac{5}{2} \end{pmatrix} \\ &\Rightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(2)} = \frac{1}{4} \begin{pmatrix} 10 - 3 \times \frac{5}{2} - 3 \times \frac{5}{2} \\ 10 - 3 \times \frac{5}{2} - 3 \times \frac{5}{2} \\ 10 - 3 \times \frac{5}{2} - 3 \times \frac{5}{2} \end{pmatrix} = \begin{pmatrix} -\frac{5}{4} \\ -\frac{5}{4} \\ -\frac{5}{4} \end{pmatrix} \Rightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(3)} = \frac{1}{4} \begin{pmatrix} 10 - 3 \times -\frac{5}{4} - 3 \times -\frac{5}{4} \\ 10 - 3 \times -\frac{5}{4} - 3 \times -\frac{5}{4} \\ 10 - 3 \times -\frac{5}{4} - 3 \times -\frac{5}{4} \end{pmatrix} = \begin{pmatrix} \frac{35}{8} \\ \frac{35}{8} \\ \frac{35}{8} \end{pmatrix} \end{aligned}$$

3.2. GAUSS-SEIDEL pour le système (5.1) :

$$\begin{aligned} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(0)} &= \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \Rightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(1)} = \frac{1}{4} \begin{pmatrix} 10 - 3 \times 0 - 3 \times 0 \\ 10 - 3 \times \frac{5}{2} - 3 \times 0 \\ 10 - 3 \times \frac{5}{2} - 3 \times \frac{5}{8} \end{pmatrix} = \begin{pmatrix} \frac{5}{2} \\ \frac{5}{8} \\ \frac{5}{32} \end{pmatrix} \\ &\Rightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(2)} = \frac{1}{4} \begin{pmatrix} 10 - 3 \times \frac{5}{8} - 3 \times \frac{5}{32} \\ 10 - 3 \times \frac{245}{128} - 3 \times \frac{5}{32} \\ 10 - 3 \times \frac{245}{128} - 3 \times \frac{485}{512} \end{pmatrix} = \begin{pmatrix} \frac{245}{128} \\ \frac{485}{512} \\ \frac{725}{2048} \end{pmatrix} \Rightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(3)} = \begin{pmatrix} \frac{12485}{8192} \\ \frac{35765}{32768} \\ \frac{70565}{131072} \end{pmatrix} \end{aligned}$$

3.3. JACOBI pour le système (5.2) :

$$\begin{aligned} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(0)} &= \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \Rightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(1)} = \frac{1}{4} \begin{pmatrix} 6 - 1 \times 0 - 1 \times 0 \\ 6 - 1 \times 0 - 1 \times 0 \\ 6 - 1 \times 0 - 1 \times 0 \end{pmatrix} = \begin{pmatrix} \frac{3}{2} \\ \frac{3}{2} \\ \frac{3}{2} \end{pmatrix} \\ &\Rightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(2)} = \frac{1}{4} \begin{pmatrix} 6 - 1 \times \frac{3}{2} - 1 \times \frac{3}{2} \\ 6 - 1 \times \frac{3}{2} - 1 \times \frac{3}{2} \\ 6 - 1 \times \frac{3}{2} - 1 \times \frac{3}{2} \end{pmatrix} = \begin{pmatrix} \frac{3}{4} \\ \frac{3}{4} \\ \frac{3}{4} \end{pmatrix} \Rightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(3)} = \frac{1}{4} \begin{pmatrix} 6 - 1 \times \frac{3}{4} - 1 \times \frac{3}{4} \\ 6 - 1 \times \frac{3}{4} - 1 \times \frac{3}{4} \\ 6 - 1 \times \frac{3}{4} - 1 \times \frac{3}{4} \end{pmatrix} = \begin{pmatrix} \frac{9}{8} \\ \frac{9}{8} \\ \frac{9}{8} \end{pmatrix} \end{aligned}$$

3.4. GAUSS-SEIDEL pour le système (5.2) :

$$\begin{aligned} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(0)} &= \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \Rightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(1)} = \frac{1}{4} \begin{pmatrix} 6 - 1 \times 0 - 1 \times 0 \\ 6 - 1 \times \frac{3}{2} - 1 \times 0 \\ 6 - 1 \times \frac{3}{2} - 1 \times \frac{9}{8} \end{pmatrix} = \begin{pmatrix} \frac{3}{2} \\ \frac{9}{8} \\ \frac{27}{32} \end{pmatrix} \\ &\Rightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(2)} = \frac{1}{4} \begin{pmatrix} 6 - 1 \times \frac{9}{8} - 1 \times \frac{27}{32} \\ 6 - 1 \times \frac{129}{128} - 1 \times \frac{27}{32} \\ 6 - 1 \times \frac{129}{128} - 1 \times \frac{531}{512} \end{pmatrix} = \begin{pmatrix} \frac{129}{128} \\ \frac{531}{512} \\ \frac{2025}{2048} \end{pmatrix} \Rightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(3)} = \frac{1}{4} \begin{pmatrix} 6 - 1 \times \frac{531}{512} - 1 \times \frac{2025}{2048} \\ 6 - 1 \times \frac{8139}{8192} - 1 \times \frac{2025}{2048} \\ 6 - 1 \times \frac{8139}{8192} - 1 \times \frac{32913}{32768} \end{pmatrix} = \begin{pmatrix} \frac{8139}{8192} \\ \frac{32913}{32768} \\ \frac{131139}{131072} \end{pmatrix} \end{aligned}$$

4. Calcul de la solution exacte à l'aide de la méthode d'élimination de GAUSS :

\* Système (5.1) :

$$\left( \begin{array}{ccc|c} 4 & 3 & 3 & 10 \\ 3 & 4 & 3 & 10 \\ 3 & 3 & 4 & 10 \end{array} \right) \xrightarrow{\substack{L_2 \leftarrow L_2 - \frac{3}{4}L_1 \\ L_3 \leftarrow L_3 - \frac{3}{4}L_1}} \left( \begin{array}{ccc|c} 4 & 3 & 3 & 10 \\ 0 & \frac{7}{4} & \frac{3}{4} & \frac{5}{2} \\ 0 & \frac{3}{4} & \frac{7}{4} & \frac{5}{2} \end{array} \right) \xrightarrow{L_3 \leftarrow L_3 - \frac{3/4}{7/4}L_2} \left( \begin{array}{ccc|c} 4 & 3 & 3 & 10 \\ 0 & \frac{7}{4} & \frac{3}{4} & \frac{5}{2} \\ 0 & 0 & \frac{10}{7} & \frac{10}{7} \end{array} \right) \implies \mathbf{x} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

\* Système (5.2) :

$$\left( \begin{array}{ccc|c} 4 & 1 & 1 & 6 \\ 1 & 4 & 1 & 6 \\ 1 & 1 & 4 & 6 \end{array} \right) \xrightarrow{\substack{L_2 \leftarrow L_2 - \frac{1}{4}L_1 \\ L_3 \leftarrow L_3 - \frac{1}{4}L_1}} \left( \begin{array}{ccc|c} 4 & 1 & 1 & 6 \\ 0 & \frac{15}{4} & \frac{3}{4} & \frac{9}{2} \\ 0 & \frac{3}{4} & \frac{15}{4} & \frac{9}{2} \end{array} \right) \xrightarrow{L_3 \leftarrow L_3 - \frac{3/4}{15/4}L_2} \left( \begin{array}{ccc|c} 4 & 1 & 1 & 6 \\ 0 & \frac{15}{4} & \frac{3}{4} & \frac{9}{2} \\ 0 & 0 & \frac{18}{5} & \frac{18}{5} \end{array} \right) \implies \mathbf{x} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$



# Chapitre 6

## Valeurs propres et vecteurs propres

### 6.1 Introduction

Soit  $\mathbb{K} = \mathbb{R}$  ou  $\mathbb{C}$ . On dit que le scalaire  $\lambda \in \mathbb{K}$  est une valeur propre de  $\mathbb{A}$  s'il existe un vecteur  $\mathbf{x} \neq \mathbf{0}$  tel que

$$\mathbb{A}\mathbf{x} = \lambda\mathbf{x}$$

où  $\mathbb{A}$  est une matrice carrée d'ordre  $n$  donnée.

On peut réécrire l'équation précédente sous la forme

$$(\mathbb{A} - \lambda\mathbb{I})\mathbf{x} = \mathbf{0}$$

qui est un système linéaire homogène de  $n$  équations. Si  $\det(\mathbb{A} - \lambda\mathbb{I}) \neq 0$  pour tout  $\lambda$ , ce système admet une et une seule solution, le vecteur  $\mathbf{x} = \mathbf{0}$ . Une solution  $\mathbf{x} \neq \mathbf{0}$  existe si et seulement si  $\det(\mathbb{A} - \lambda\mathbb{I}) = 0$ .

- ★ On appelle POLYNÔME CARACTÉRISTIQUE DE LA MATRICE  $\mathbb{A}$  le polynôme défini par

$$p_{\mathbb{A}}(\lambda) = \det(\mathbb{A} - \lambda\mathbb{I}) = a_0 + a_1\lambda + a_2\lambda^2 + \cdots + a_n\lambda^n.$$

Dans  $\mathbb{C}$ , tout polynôme admet exactement  $n$  racines (comptées avec leur multiplicité).

Dans  $\mathbb{R}$ , tout polynôme admet au plus  $n$  racines (comptées avec leur multiplicité).

- ★ On appelle VALEUR PROPRE DE  $\mathbb{A}$  tout élément  $\lambda \in \mathbb{K}$  tel que  $p(\lambda) = 0$ .

La multiplicité de la valeur propre est dite "multiplicité algébrique".

Dans  $\mathbb{C}$ , toute matrice carrée d'ordre  $n$  admet exactement  $n$  valeurs propres (distinctes ou confondues).

Dans  $\mathbb{R}$ , toute matrice carrée d'ordre  $n$  admet donc au plus  $n$  valeurs propres (distinctes ou confondues).

- ★ On appelle SPECTRE DE  $\mathbb{A}$  l'ensemble de ses valeurs propres et on le note  $\sigma(\mathbb{A})$ .

- ★ On appelle RAYON SPECTRALE DE  $\mathbb{A}$  la valeur propre de module maximale et on le note  $\rho(\mathbb{A})$ .

- ★ On appelle VECTEUR PROPRE DE  $\mathbb{A}$  ASSOCIÉ À LA VALEUR PROPRE  $\lambda$  tout vecteur  $\mathbf{x} \neq \mathbf{0}$  tel que  $(\mathbb{A} - \lambda\mathbb{I})\mathbf{x} = \mathbf{0}$ .

- ★ L'ensemble des VECTEURS PROPRES DE  $\mathbb{A}$  ASSOCIÉS À LA VALEUR PROPRE  $\lambda$  engendre un espace vectoriel. La dimension de cet espace vectoriel est dite "multiplicité géométrique" et elle toujours inférieure ou égale à la multiplicité algébrique de la valeur propre correspondante.

On peut démontrer que

1.  $\det(\mathbb{A}) = \prod_{i=1}^n \lambda_i$ , (donc  $\det(\mathbb{A}) = 0$ ssi il existe une valeur propre nulle) ;
2.  $\text{tr}(\mathbb{A}) = \sum_{i=1}^n \lambda_i$  ;
3.  $\sigma(\mathbb{A}^T) = \sigma(\mathbb{A})$  et  $\sigma(\mathbb{A}^H) = \sigma(\mathbb{A})$  ;
4.  $\lambda$  est une valeur propre de  $\mathbb{A} \in \mathbb{C}^{n \times n} \iff \bar{\lambda}$  est une valeur propre de  $\mathbb{A}^H$ .

On dit que deux matrices  $\mathbb{A}$  et  $\mathbb{B}$  carrées d'ordre  $n$  sont semblables s'il existe une matrice  $\mathbb{P}$  carrée d'ordre  $n$  inversible telle que  $\mathbb{A} = \mathbb{P}^{-1}\mathbb{B}\mathbb{P}$ . On peut démontrer que

1.  $p_{\mathbb{A}}(\lambda) = p_{\mathbb{B}}(\lambda)$  ;
2.  $\sigma(\mathbb{A}) = \sigma(\mathbb{B})$ .

Une matrice carrée  $\mathbb{A}$  d'ordre  $n$  est diagonalisable si elle est semblable à une matrice diagonale. On peut démontrer que

1. si le polynôme caractéristique a exactement  $n$  racines distinctes deux à deux alors  $\mathbb{A}$  est diagonalisable et  $\mathbb{A} = \mathbb{P}^{-1}\mathbb{D}\mathbb{P}$  avec  $\mathbb{D} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$  et les colonnes de  $\mathbb{P}$  sont les vecteurs propres de  $\mathbb{A}$  ;
2.  $\mathbb{A}^P = \mathbb{P}^{-1}\mathbb{D}^P\mathbb{P}$  et  $\mathbb{D}^P = \text{diag}(\lambda_1^P, \lambda_2^P, \dots, \lambda_n^P)$  ;
3. si la "multiplicité géométrique" de l'espace vectoriel associé à une valeur propre est strictement inférieur à la "multiplicité algébrique" de cette valeur propre, la matrice n'est pas diagonalisable ;

4. si  $\mathbb{A}$  est orthogonale alors elle est diagonalisable sur  $\mathbb{C}$ .

Par conséquent, une matrice peut être diagonalisable dans  $\mathbb{C}$  mais pas dans  $\mathbb{R}$ .

Une matrice carrée  $\mathbb{A}$  d'ordre  $n$  est trigonalisable si elle est semblable à une matrice triangulaire. On peut démontrer que toute matrice est trigonalisable sur  $\mathbb{C}$  et l'on a le résultat suivant :

### Proposition 6.1 (Décomposition de Schur)

Pour toute matrice  $\mathbb{A} \in \mathbb{C}^{n \times n}$  il existe une matrice  $\mathbb{U}$  carrées d'ordre  $n$  unitaire telle que  $\mathbb{T} = \mathbb{U}^{-1}\mathbb{A}\mathbb{U} = \mathbb{U}^H\mathbb{A}\mathbb{U}$  avec

$$\mathbb{T} = \begin{pmatrix} \lambda_1 & t_{12} & \dots & t_{1n} \\ 0 & \lambda_2 & t_{12} & t_{2n} \\ \vdots & & \ddots & \vdots \\ 0 & \dots & 0 & \lambda_n \end{pmatrix}$$

ayant noté  $\lambda_i$  les valeurs propres de  $\mathbb{A}$ . Les matrices  $\mathbb{U}$  et  $\mathbb{T}$  ne sont pas forcément uniques.

On peut démontrer que

1. si  $\mathbb{A}$  est hermitienne alors  $\mathbb{T}$  est toujours une matrice diagonale et les colonnes de  $\mathbb{U}$  sont les vecteurs propres de  $\mathbb{A}$  ;
2. si de plus  $\mathbb{A}$  est normale alors on a la décomposition spectrale  $\mathbb{A} = \mathbb{U}\mathbb{T}\mathbb{U}^H = \sum_{i=1}^n \lambda_i \mathbf{u}_i \mathbf{u}_i^H$

### EXEMPLE

Considérons la matrice

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

#### \* Calcul des valeurs propres

Le polynôme caractéristique de  $\mathbb{A}$  est

$$\begin{aligned} p(\lambda) &= \det(\mathbb{A} - \lambda \mathbb{I}) = \det \begin{pmatrix} 1-\lambda & -1 & 0 \\ -1 & 2-\lambda & -1 \\ 0 & -1 & 1-\lambda \end{pmatrix} \\ &= (1-\lambda) \det \begin{pmatrix} 2-\lambda & -1 \\ -1 & 1-\lambda \end{pmatrix} - (-1) \det \begin{pmatrix} -1 & -1 \\ 0 & 1-\lambda \end{pmatrix} \\ &= (1-\lambda)((2-\lambda)(1-\lambda)-1) - (1-\lambda) = (1-\lambda)((2-\lambda)(1-\lambda)-2) = (1-\lambda)(-3\lambda+\lambda^2) = \lambda(1-\lambda)(\lambda-3) \end{aligned}$$

Nous avons trouvé 3 valeurs propres :

$$\lambda_1 = 0 \quad < \quad \lambda_2 = 1 \quad < \quad \lambda_3 = 3.$$

#### \* Calcul des vecteurs propres

\* Calcul des vecteurs propres associés à la valeur propre  $\lambda_1$ .

On cherche  $\mathbf{x}$  tel que

$$(\mathbb{A} - \lambda_1 \mathbb{I})\mathbf{x} = \mathbf{0} \quad \text{c'est-à-dire} \quad \begin{pmatrix} 1-\lambda_1 & -1 & 0 \\ -1 & 2-\lambda_1 & -1 \\ 0 & -1 & 1-\lambda_1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

En utilisant la méthode de Gauss on a

$$\begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix} \xrightarrow[\text{Étape 1}]{\begin{array}{l} L_2 \leftarrow L_2 + L_1 \\ L_3 \leftarrow L_3 - L_1 \end{array}} \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \end{pmatrix} \xrightarrow[\text{Étape 2}]{L_3 \leftarrow L_3 + L_2} \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{pmatrix}$$

On obtient le système linéaire triangulaire supérieure

$$\begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

donc  $x_3 = \kappa \in \mathbb{R}$ ,  $x_2 = x_3 = \kappa$  et  $x_1 = x_2 = \kappa$  donc

$$\mathbf{x} = \kappa \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

Il est usuel de choisir  $\kappa$  de sorte à ce que  $\mathbf{x}$  soit normalisé, donc  $\kappa = 1/\|(1, 1, 1)\| = 1/\sqrt{1^2 + 1^2 + 1^2} = 1/\sqrt{3}$  et on pose

$$\mathbf{x}_1 = \begin{pmatrix} \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{pmatrix}.$$

\* Calcul des vecteurs propres associés à la valeur propre  $\lambda_2$ .

On cherche  $\mathbf{x}$  tel que

$$(\mathbb{A} - \lambda_2 \mathbb{I})\mathbf{x} = \mathbf{0} \quad \text{c'est-à-dire} \quad \begin{pmatrix} 1 - \lambda_2 & -1 & 0 \\ -1 & 2 - \lambda_2 & -1 \\ 0 & -1 & 1 - \lambda_2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

En utilisant la méthode de Gauss on a

$$\begin{pmatrix} 0 & -1 & 0 \\ -1 & 1 & -1 \\ 0 & -1 & 0 \end{pmatrix} \xrightarrow[\text{Étape 1}]{L_2 \leftarrow L_2 - L_1} \begin{pmatrix} -1 & 1 & -1 \\ 0 & -1 & 0 \\ 0 & -1 & 0 \end{pmatrix} \xrightarrow[\text{Étape 2}]{L_3 \leftarrow L_3 - L_2} \begin{pmatrix} -1 & 1 & -1 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

On obtient le système linéaire triangulaire supérieure

$$\begin{pmatrix} -1 & 1 & -1 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

donc  $x_3 = \kappa \in \mathbb{R}$ ,  $x_2 = 0$  et  $x_1 = x_2 - x_3 = -\kappa$  donc

$$\mathbf{x} = \kappa \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}.$$

Il est usuel de choisir  $\kappa$  de sorte à ce que  $\mathbf{x}$  soit normalisé, donc  $\kappa = 1/\|(-1, 0, 1)^T\| = 1/\sqrt{(-1)^2 + 0^2 + 1^2} = 1/\sqrt{2}$  et on pose

$$\mathbf{x}_2 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ -\frac{1}{\sqrt{2}} \end{pmatrix}.$$

\* Calcul des vecteurs propres associés à la valeur propre  $\lambda_3$ .

On cherche  $\mathbf{x}$  tel que

$$(\mathbb{A} - \lambda_3 \mathbb{I})\mathbf{x} = \mathbf{0} \quad \text{c'est-à-dire} \quad \begin{pmatrix} 1 - \lambda_3 & -1 & 0 \\ -1 & 2 - \lambda_3 & -1 \\ 0 & -1 & 1 - \lambda_3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

En utilisant la méthode de Gauss on a

$$\begin{pmatrix} -2 & -1 & 0 \\ -1 & -1 & -1 \\ 0 & -1 & -2 \end{pmatrix} \xrightarrow[\text{Étape 1}]{L_2 \leftarrow L_2 - \frac{1}{2}L_1} \begin{pmatrix} -2 & -1 & 0 \\ 0 & -\frac{1}{2} & -1 \\ 0 & -1 & -2 \end{pmatrix} \xrightarrow[\text{Étape 2}]{L_3 \leftarrow L_3 - 2L_2} \begin{pmatrix} -2 & -1 & 0 \\ 0 & -\frac{1}{2} & -1 \\ 0 & 0 & 0 \end{pmatrix}$$

On obtient le système linéaire triangulaire supérieure

$$\begin{pmatrix} -2 & -1 & 0 \\ 0 & -\frac{1}{2} & -1 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

donc  $x_3 = \kappa \in \mathbb{R}$ ,  $x_2 = -2x_3 = -2\kappa$  et  $x_1 = -x_2/2 = \kappa$  donc

$$\mathbf{x} = \kappa \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}.$$

Il est usuel de choisir  $\kappa$  de sorte à ce que  $\mathbf{x}$  soit normalisé, donc  $\kappa = 1/\|(1, 2, 1)^T\| = 1/\sqrt{1^2 + 2^2 + 1^2} = 1/\sqrt{6}$  et on pose

$$\mathbf{x}_3 = \begin{pmatrix} \frac{1}{\sqrt{6}} \\ -\frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{pmatrix}.$$

On peut alors écrire les valeurs propres et les vecteurs propres dans deux matrices

$$\mathbb{D} = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{pmatrix} \quad \text{et} \quad \mathbb{P} = (\mathbf{x}_1 \quad \mathbf{x}_2 \quad \mathbf{x}_3) = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \end{pmatrix}$$

et vérifier que  $\mathbb{A} = \mathbb{P}\mathbb{D}\mathbb{P}^{-1}$ , c'est-à-dire que  $\mathbb{A}\mathbb{P} = \mathbb{P}\mathbb{D}$  :

$$\begin{aligned} \mathbb{A}\mathbb{P} &= \begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{\sqrt{2}} & -\frac{3}{\sqrt{6}} \\ 0 & 0 & -\frac{3}{\sqrt{6}} \\ 0 & -\frac{1}{\sqrt{2}} & \frac{3}{\sqrt{6}} \end{pmatrix} \\ \mathbb{P}\mathbb{D} &= \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{\sqrt{2}} & \frac{3}{\sqrt{6}} \\ 0 & 0 & -\frac{3}{\sqrt{6}} \\ 0 & -\frac{1}{\sqrt{2}} & \frac{3}{\sqrt{6}} \end{pmatrix} \end{aligned}$$

## 6.2 Localisation des valeurs propres

Soit  $\mathbb{A}$  une matrice carrée d'ordre  $n$ .

Une première estimation de la localisation du spectre d'une matrice dans le plan complexe est donnée par

$$|\lambda| \leq \|\mathbb{A}\| \quad \forall \lambda \in \sigma(\mathbb{A})$$

pour toute norme  $\|\cdot\|$  consistante. Cette estimation dit que toutes les valeurs propres appartiennent au cercle de rayon  $\|\mathbb{A}\|$  centré dans l'origine du plan complexe :

$$\sigma(\mathbb{A}) \subset \mathcal{O} = \{z \in \mathbb{C} \mid |z| \leq \|\mathbb{A}\|\}.$$

Parmi les normes les plus utilisées nous avons

$$\begin{aligned} \|\mathbb{A}\|_1 &= \max_{j=1 \dots n} \sum_{i=1}^n |a_{ij}| = \|\mathbb{A}^T\|_\infty \\ \|\mathbb{A}\|_\infty &= \max_{i=1 \dots n} \sum_{j=1}^n |a_{ij}| = \|\mathbb{A}^T\|_1 \end{aligned}$$

Une autre estimation est donnée par les disques de GERSHGORIN. Les disques de GERSHGORIN  $\mathcal{R}_i$  et  $\mathcal{C}_j$  associés à la  $i$ -ème ligne et à la  $j$ -ème colonne sont respectivement définis par

$$\mathcal{R}_i = \left\{ z \in \mathbb{C} \mid |z - a_{ii}| \leq \sum_{j=1, j \neq i}^n |a_{ij}| \right\}, \quad \mathcal{C}_j = \left\{ z \in \mathbb{C} \mid |z - a_{jj}| \leq \sum_{i=1, i \neq j}^n |a_{ij}| \right\}.$$

Les disques de GERSHGORIN peuvent servir à localiser les valeurs propres d'une matrice, comme le montre la proposition suivante

### Proposition 6.2

Toutes les valeurs propres d'une matrice  $\mathbb{A} \in \mathbb{C}^{n \times n}$  appartiennent à la région du plan complexe définie par l'intersection des

deux régions constituées respectivement de la réunion des disques des lignes et des disques des colonnes :

$$\sigma(\mathbb{A}) \subset \underbrace{\left( \bigcup_{i=1}^n \mathcal{R}_i \right)}_{\mathcal{S}_{\mathcal{R}}} \cap \underbrace{\left( \bigcup_{j=1}^n \mathcal{C}_j \right)}_{\mathcal{S}_{\mathcal{C}}}.$$

Si de plus  $m$  disques des lignes (ou des colonnes),  $1 \leq m \leq n$ , sont disjoints de la réunion des  $n - m$  autres disques, alors leur réunion contient exactement  $m$  valeurs propres.

Rien n'assure qu'un disque contienne des valeurs propres, à moins qu'il ne soit isolé des autres.

Remarquer qu'on peut déduire que toutes les valeurs propres d'une matrice à diagonale strictement dominante sont non nulles.

#### EXEMPLE

Considérons la matrice

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

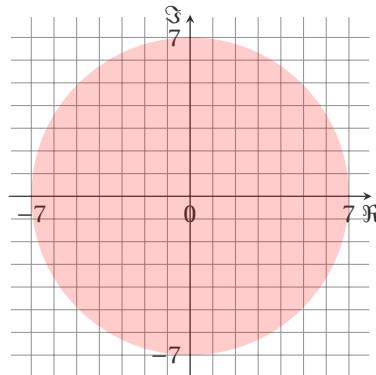
Nous avons les estimations suivantes :

- Si on considère les normes  $\|\cdot\|_1$  et  $\|\cdot\|_\infty$  nous avons

$$\begin{aligned} \|\mathbb{A}\|_1 &= \max_{j=1 \dots n} \sum_{i=1}^n |a_{ij}| = \max_{j=1 \dots n} \{ |3| + |-1| + |0|; |2| + |2| + |1|; |3| + |-1| + |3| \} = \max_{j=1 \dots n} \{ 4; 5; 7 \} = 7 \\ \|\mathbb{A}\|_\infty &= \max_{i=1 \dots n} \sum_{j=1}^n |a_{ij}| = \max_{i=1 \dots n} \{ |3| + |2| + |3|; |-1| + |2| + |-1|; |0| + |1| + |3| \} = \max_{i=1 \dots n} \{ 8; 4; 4 \} = 8 \end{aligned}$$

donc toutes les valeurs propres appartiennent au cercle de rayon 7 centré dans l'origine du plan complexe :

$$\sigma(\mathbb{A}) \subset \mathcal{O} = \{ z \in \mathbb{C} \mid |z| \leq 7 \}.$$

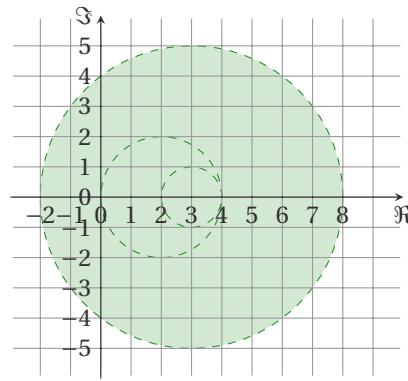


- Disques des lignes :

$$\begin{aligned} \mathcal{R}_1 &= \left\{ z \in \mathbb{C} \mid |z - a_{11}| \leq \sum_{\substack{j=1, \\ j \neq 1}}^n |a_{1j}| \right\} = \{ z \in \mathbb{C} \mid |z - 3| \leq |2| + |3| \} = \{ z \in \mathbb{C} \mid |z - 3| \leq 5 \}, \\ \mathcal{R}_2 &= \left\{ z \in \mathbb{C} \mid |z - a_{22}| \leq \sum_{\substack{j=1, \\ j \neq 2}}^n |a_{2j}| \right\} = \{ z \in \mathbb{C} \mid |z - 2| \leq |-1| + |-1| \} = \{ z \in \mathbb{C} \mid |z - 2| \leq 2 \}, \\ \mathcal{R}_3 &= \left\{ z \in \mathbb{C} \mid |z - a_{33}| \leq \sum_{\substack{j=1, \\ j \neq 3}}^n |a_{3j}| \right\} = \{ z \in \mathbb{C} \mid |z - 3| \leq |0| + |1| \} = \{ z \in \mathbb{C} \mid |z - 3| \leq 1 \}. \end{aligned}$$

Toutes les valeurs propres appartiennent à la réunion des disques des lignes :

$$\sigma(\mathbb{A}) \subset \mathcal{S}_{\mathcal{R}} = \mathcal{R}_1 \cup \mathcal{R}_2 \cup \mathcal{R}_3 = \mathcal{R}_1.$$

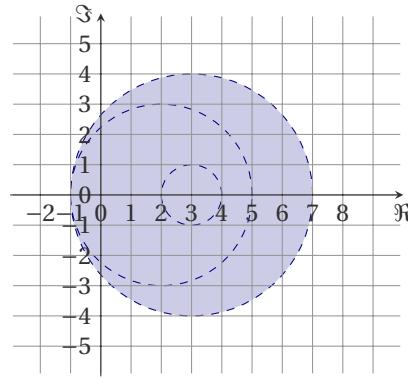


3. Disques des colonnes :

$$\begin{aligned}\mathcal{C}_1 &= \left\{ z \in \mathbb{C} \mid |z - a_{11}| \leq \sum_{\substack{i=1, \\ i \neq 1}}^n |a_{i1}| \right\} = \{ z \in \mathbb{C} \mid |z - 3| \leq |-1| + |0| \} = \{ z \in \mathbb{C} \mid |z - 3| \leq 1 \}, \\ \mathcal{C}_2 &= \left\{ z \in \mathbb{C} \mid |z - a_{22}| \leq \sum_{\substack{i=1, \\ i \neq 2}}^n |a_{i2}| \right\} = \{ z \in \mathbb{C} \mid |z - 2| \leq |2| + |1| \} = \{ z \in \mathbb{C} \mid |z - 2| \leq 3 \}, \\ \mathcal{C}_3 &= \left\{ z \in \mathbb{C} \mid |z - a_{33}| \leq \sum_{\substack{i=1, \\ i \neq 3}}^n |a_{i3}| \right\} = \{ z \in \mathbb{C} \mid |z - 3| \leq |3| + |-1| \} = \{ z \in \mathbb{C} \mid |z - 3| \leq 4 \}.\end{aligned}$$

Toutes les valeurs propres appartiennent à la réunion des disques des colonnes :

$$\sigma(\mathbb{A}) \subset \mathcal{S}_{\mathcal{C}} = \mathcal{C}_1 \cup \mathcal{C}_2 \cup \mathcal{C}_3 = \mathcal{C}_3.$$



4. Toutes les valeurs propres appartiennent à l'intersection de ces trois régions :

$$\sigma(\mathbb{A}) \subset \mathcal{S}_{\mathcal{C}} \cap \mathcal{S}_{\mathcal{R}} \cap \mathcal{S}_{\mathcal{C}} = \mathcal{C}_3$$

En effet, on a

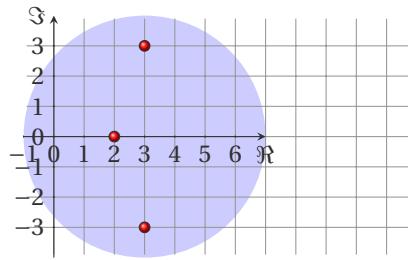
$$p_{\mathbb{A}}(\lambda) = -\lambda^3 + 8\lambda^2 - 24\lambda + 24 = -(\lambda - 2)(\lambda^2 - 6\lambda + 12)$$

donc

$$\lambda_1 = 2$$

$$\lambda_2 = 3 + i\sqrt{3}$$

$$\lambda_3 = \overline{\lambda_2} = 3 - i\sqrt{3}.$$



### EXEMPLE

Considérons la matrice

$$\mathbb{A} = \begin{pmatrix} 10 & 2 & 3 \\ -1 & 2 & -1 \\ 0 & 1 & 3 \end{pmatrix}.$$

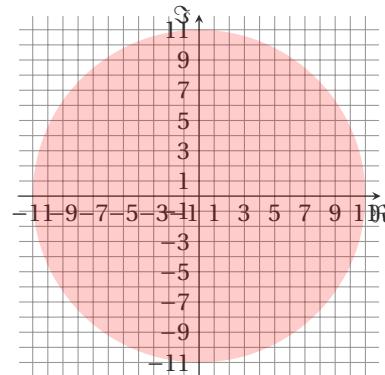
Nous avons les estimations suivantes :

- Si on considère les normes  $\|\cdot\|_1$  et  $\|\cdot\|_\infty$  nous avons

$$\begin{aligned}\|\mathbb{A}\|_1 &= \max_{j=1 \dots n} \sum_{i=1}^n |a_{ij}| = \max_{j=1 \dots n} \{ |10| + |-1| + |0|; |2| + |2| + |1|; |3| + |-1| + |3| \} = \max_{j=1 \dots n} \{ 11; 5; 7 \} = 11 \\ \|\mathbb{A}\|_\infty &= \max_{i=1 \dots n} \sum_{j=1}^n |a_{ij}| = \max_{i=1 \dots n} \{ |10| + |2| + |3|; |-1| + |2| + |-1|; |0| + |1| + |3| \} = \max_{i=1 \dots n} \{ 15; 4; 4 \} = 15\end{aligned}$$

donc toutes les valeurs propres appartiennent au cercle de rayon 11 centré dans l'origine du plan complexe :

$$\sigma(\mathbb{A}) \subset \mathcal{O} = \{ z \in \mathbb{C} \mid |z| \leq 11 \}.$$



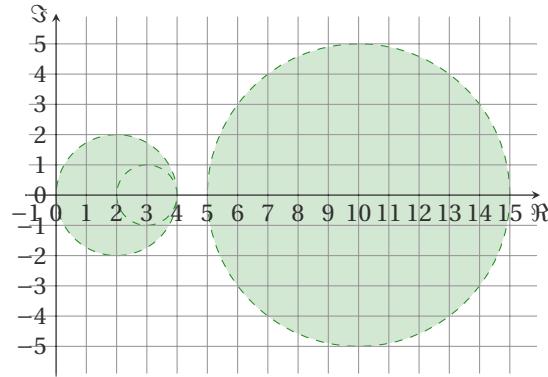
- Disques des lignes :

$$\begin{aligned}\mathcal{R}_1 &= \left\{ z \in \mathbb{C} \mid |z - a_{11}| \leq \sum_{\substack{j=1, \\ j \neq 1}}^n |a_{1j}| \right\} = \{ z \in \mathbb{C} \mid |z - 10| \leq |2| + |3| \} = \{ z \in \mathbb{C} \mid |z - 10| \leq 5 \}, \\ \mathcal{R}_2 &= \left\{ z \in \mathbb{C} \mid |z - a_{22}| \leq \sum_{\substack{j=1, \\ j \neq 2}}^n |a_{2j}| \right\} = \{ z \in \mathbb{C} \mid |z - 2| \leq |-1| + |-1| \} = \{ z \in \mathbb{C} \mid |z - 2| \leq 2 \}, \\ \mathcal{R}_3 &= \left\{ z \in \mathbb{C} \mid |z - a_{33}| \leq \sum_{\substack{j=1, \\ j \neq 3}}^n |a_{3j}| \right\} = \{ z \in \mathbb{C} \mid |z - 3| \leq |0| + |1| \} = \{ z \in \mathbb{C} \mid |z - 3| \leq 1 \}.\end{aligned}$$

Toutes les valeurs propres appartiennent à la réunion des disques des lignes :

$$\sigma(\mathbb{A}) \subset \mathcal{S}_{\mathcal{R}} = \mathcal{R}_1 \cup \mathcal{R}_2 \cup \mathcal{R}_3.$$

De plus, comme le disque  $\mathcal{R}_1$  est disjoint de la réunion  $\mathcal{R}_2 \cup \mathcal{R}_3$ , une et une seule valeur propre est contenue dans  $\mathcal{R}_1$  et les deux autres valeurs propres appartiennent à  $\mathcal{R}_2 \cup \mathcal{R}_3 = \mathcal{R}_2$ .



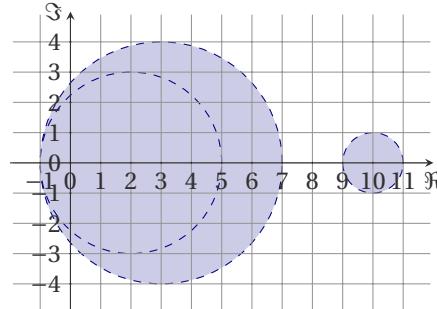
3. Disques des colonnes :

$$\begin{aligned}\mathcal{C}_1 &= \left\{ z \in \mathbb{C} \mid |z - a_{11}| \leq \sum_{\substack{i=1, \\ i \neq 1}}^n |a_{i1}| \right\} = \{z \in \mathbb{C} \mid |z - 10| \leq |-1| + |0|\} = \{z \in \mathbb{C} \mid |z - 10| \leq 1\}, \\ \mathcal{C}_2 &= \left\{ z \in \mathbb{C} \mid |z - a_{22}| \leq \sum_{\substack{i=1, \\ i \neq 2}}^n |a_{i2}| \right\} = \{z \in \mathbb{C} \mid |z - 2| \leq |2| + |1|\} = \{z \in \mathbb{C} \mid |z - 2| \leq 3\}, \\ \mathcal{C}_3 &= \left\{ z \in \mathbb{C} \mid |z - a_{33}| \leq \sum_{\substack{i=1, \\ i \neq 3}}^n |a_{i3}| \right\} = \{z \in \mathbb{C} \mid |z - 3| \leq |3| + |-1|\} = \{z \in \mathbb{C} \mid |z - 3| \leq 4\}.\end{aligned}$$

Toutes les valeurs propres appartiennent à la réunion des disques des colonnes :

$$\sigma(\mathbb{A}) \subset \mathcal{S}_{\mathcal{C}} = \mathcal{C}_1 \cup \mathcal{C}_2 \cup \mathcal{C}_3.$$

De plus, comme le disque  $\mathcal{C}_1$  est disjoint de la réunion  $\mathcal{C}_2 \cup \mathcal{C}_3$ , une et une seule valeur propre est contenue dans  $\mathcal{C}_1$  et les deux autres valeurs propres appartiennent à  $\mathcal{C}_2 \cup \mathcal{C}_3 = \mathcal{C}_3$ .



4. Toutes les valeurs propres appartiennent à l'intersection de ces trois régions :

$$\sigma(\mathbb{A}) \subset \mathcal{S}_{\mathcal{O}} \cap \mathcal{S}_{\mathcal{R}} \cap \mathcal{S}_{\mathcal{C}}$$

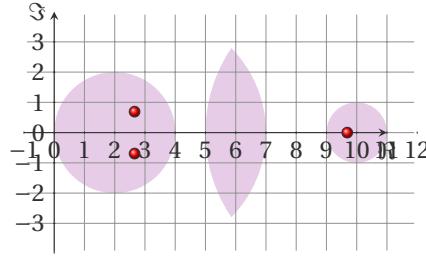
En effet, on a

$$\lambda_1 \approx 9.6876$$

$$\lambda_2 \approx 2.6562 + 0.6928i$$

$$\lambda_3 = \overline{\lambda_2} \approx 2.6562 - 0.6928i.$$

```
A=[10 2 3; -1 2 -1; 0 1 3]
eig(A)
```



## 6.3 Approximation

montrer la méthode QR

## 6.4 Décomposition en valeurs singulières

Soit  $\mathbb{A} \in \mathbb{R}^{n \times p}$  une matrice rectangulaire. Un théorème démontré officiellement en 1936 par C. ECKART et G. YOUNG affirme que toute matrice rectangulaire  $\mathbb{A}$  se décompose sous la forme

$$\mathbb{A} = \mathbb{U}\mathbb{S}\mathbb{V}^T$$

avec  $\mathbb{U} \in \mathbb{R}^{n \times n}$  et  $\mathbb{V} \in \mathbb{R}^{p \times p}$  des matrices orthogonales (*i.e.*  $\mathbb{U}^{-1} = \mathbb{U}^T$  et  $\mathbb{V}^{-1} = \mathbb{V}^T$ ) et  $\mathbb{S} \in \mathbb{R}^{n \times p}$  une matrice diagonale qui contient les  $r$  valeurs singulières de  $\mathbb{A}$ ,  $r = \min\{n, p\}$ ,  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r \geq 0$ . Ce qui est remarquable, c'est que n'importe quelle matrice admet une telle décomposition alors que la décomposition en valeurs propres (la diagonalisation d'une matrice) n'est pas toujours possible.

Notons  $\mathbf{u}_i$  et  $\mathbf{v}_i$  les vecteurs colonne des matrices  $\mathbb{U}$  et  $\mathbb{V}$ . La décomposition s'écrit alors

$$\begin{aligned} \mathbb{A} &= \mathbb{U}\mathbb{S}\mathbb{V}^T = \underbrace{\begin{pmatrix} \mathbf{u}_1 & \dots & \mathbf{u}_r & \mathbf{u}_{r+1} & \dots & \mathbf{u}_n \end{pmatrix}}_{\in \mathbb{R}^{n \times n}} \underbrace{\begin{pmatrix} \sigma_1 & & & & & \\ & \ddots & & & & \\ & & \sigma_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix}}_{\in \mathbb{R}^{n \times p}} \underbrace{\begin{pmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_r^T \\ \mathbf{v}_{r+1}^T \\ \vdots \\ \mathbf{v}_p^T \end{pmatrix}}_{\in \mathbb{R}^{p \times p}} \\ &= \underbrace{\begin{pmatrix} \mathbf{u}_1 & \dots & \mathbf{u}_r \end{pmatrix}}_{\in \mathbb{R}^{n \times r}} \underbrace{\begin{pmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_r & \\ & & & \mathbf{v}_r^T \end{pmatrix}}_{\in \mathbb{R}^{r \times p}} = \sum_{i=1}^r \sigma_i \underbrace{\mathbf{u}_i \times \mathbf{v}_i^T}_{\in \mathbb{R}^{r \times r}} \end{aligned}$$

Pour calculer ces trois matrices on remarque que  $\mathbb{A} = \mathbb{U}\mathbb{S}\mathbb{V}^T = \mathbb{U}\mathbb{S}\mathbb{V}^{-1}$  et  $\mathbb{A}^T = \mathbb{V}\mathbb{S}\mathbb{U}^T = \mathbb{V}\mathbb{S}\mathbb{U}^{-1}$  donc

$$\begin{aligned} \mathbb{A}^T \mathbf{u}_i &= \sigma_i \mathbf{v}_i, & \text{pour } i = 1, \dots, r \\ \mathbb{A} \mathbf{v}_i &= \sigma_i \mathbf{u}_i, & \text{pour } i = 1, \dots, r \end{aligned}$$

ainsi

$$\begin{aligned} \mathbb{A} \mathbb{A}^T \mathbf{u}_i &= \sigma_i \mathbb{A} \mathbf{v}_i = \sigma_i^2 \mathbf{u}_i, & \text{pour } i = 1, \dots, r \\ \mathbb{A}^T \mathbb{A} \mathbf{v}_i &= \sigma_i \mathbb{A}^T \mathbf{u}_i = \sigma_i^2 \mathbf{v}_i, & \text{pour } i = 1, \dots, r \end{aligned}$$

ainsi les  $\sigma_i^2$  sont les valeurs propres de la matrice  $\mathbb{A} \mathbb{A}^T$  et les  $\mathbf{u}_i$  les vecteurs propres associés mais aussi les  $\sigma_i^2$  sont les valeurs propres de la matrice  $\mathbb{A}^T \mathbb{A}$  et les  $\mathbf{v}_i$  les vecteurs propres associés (attention, étant des valeurs propres, ils ne sont pas définis de façon unique).

On peut exploiter cette décomposition pour faire des économies de mémoire. En effet, pour stocker la matrice  $\mathbb{A}$  nous avons besoin de  $n \times p$  valeurs, pour stocker la décomposition SVD nous avons besoin de  $n \times r + r + r \times p > n \times p$  valeurs. Cependant, si nous approchons  $\mathbb{A}$  en ne gardant que les premiers termes de la somme (sachant que les derniers termes sont

multipliés par des  $\sigma_i$  plus petits, voire nuls)

$$\tilde{\mathbb{A}} = \sum_{i=1}^s \sigma_i \underbrace{\mathbf{u}_i \times \mathbf{v}_i^T}_{\in \mathbb{R}^{n \times p}}, \quad \text{où } s < r$$

nous n'avons plus besoin que de  $n \times s + s + s \times p$  valeurs. Cela signifie que pour tout  $s < np/(n+p+1)$  on fait des économies de stockage.

Expliquer l'algo page rank de google



## Exercices



### ★ Exercice 6.1 (Valeurs et vecteurs propres)

Soit la matrice

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

Vérifier numériquement avec les fonctions déjà implémentées dans Octave que le déterminant, le rayon spectral, les valeurs propres et vecteurs propres sont ceux qu'on a calculé analytiquement en cours.

Vérifier numériquement que  $\mathbb{V}^{-1}\mathbb{A}\mathbb{V} = \mathbb{D}$  où  $\mathbb{D}$  est la matrice diagonale qui contient les valeurs propres et  $\mathbb{V}$  la matrice dont les colonnes sont les vecteurs propres associés.

Utiliser la factorisation QR pour approcher les valeurs propres de  $\mathbb{A}$  avec une erreur inférieure à  $10^{-14}$ .

#### Correction

```
A = [1 -1 0; -1 2 -1; 0 -1 1]
```

```
detA=det(A)
Lambda=eig(A)
rho = max(abs(Lambda))
[V,D] = eig(A)
erreur=D-inv(V)*A*V
norm(erreur)

T=A;
niter=0;
test=norm(tril(A,-1),inf)+1;
while test>=10.e-14 & niter<=10^5
    [Q,R]=qr(T);
    T=R*Q;
    niter+=1;
    test=norm(tril(T,-1),inf);
end
D=diag(T)
niter
test
```

### ★ Exercice 6.2 (Valeurs et vecteurs propres)

Soit la matrice

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

Vérifier numériquement avec les fonctions déjà implémentées dans Octave que le déterminant, le rayon spectral, les valeurs propres et vecteurs propres sont ceux qu'on a calculé analytiquement en TD.

Vérifier numériquement que  $\mathbb{V}^{-1}\mathbb{A}\mathbb{V} = \mathbb{D}$  où  $\mathbb{D}$  est la matrice diagonale qui contient les valeurs propres et  $\mathbb{V}$  la matrice dont les colonnes sont les vecteurs propres associés.

Utiliser la factorisation QR pour approcher les valeurs propres de  $\mathbb{A}$  avec une erreur inférieure à  $10^{-14}$ .

#### Correction

```
A = [1 -1 0; -1 2 -1; 0 -1 1]
```

```
detA=det(A)
Lambda=eig(A)
rho = max(abs(Lambda))
[V,D] = eig(A)
erreur=D-inv(V)*A*V
norm(erreur)

T=A;
niter=0;
test=norm(tril(A,-1),inf)+1;
while test>=10.e-14 & niter<=10^5
```

```
[Q,R]=qr(T);
T=R*Q;
niter+=1;
test=norm(tril(T,-1),inf);
end
D=diag(T)
niter
test
```

### ★ Exercice 6.3 (Valeurs singulières)

Soit

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

Calculer analytiquement et vérifier numériquement sa décomposition SVD.

#### Correction

$\mathbb{A} \in \mathbb{R}^{n \times p}$  avec  $n = 2$  et  $p = 3$  donc  $r = 2$ .

Pour calculer la décomposition SVD nous allons calculer les valeurs et vecteurs propres des matrices  $\mathbb{A}\mathbb{A}^T$  et  $\mathbb{A}^T\mathbb{A}$ .

$$\mathbb{A}\mathbb{A}^T = \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix} \quad \mathbb{A}^T\mathbb{A} = \begin{pmatrix} 5 & 4 & 0 \\ 4 & 5 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Valeurs propres :

$$\lambda_1 = 9 > \lambda_2 = 1$$

$$\lambda_1 = 2 > \lambda_2 = 1 > \lambda_3 = 0$$

Vecteurs propres unitaires :

$$\mathbb{U} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

$$\mathbb{V} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & \sqrt{2} \end{pmatrix}$$

Donc

$$\begin{aligned} \mathbb{A} = \mathbb{U} \mathbb{S} \mathbb{V}^T &= \underbrace{\begin{pmatrix} \mathbf{u}_1 & \dots & \mathbf{u}_r & \mathbf{u}_{r+1} & \dots & \mathbf{u}_n \end{pmatrix}}_{\in \mathbb{R}^{n \times n}} \underbrace{\begin{pmatrix} \sigma_1 & & & & & \\ & \ddots & & & & \\ & & \sigma_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix}}_{\in \mathbb{R}^{n \times p}} \underbrace{\begin{pmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_r^T \\ \mathbf{v}_{r+1}^T \\ \vdots \\ \mathbf{v}_p^T \end{pmatrix}}_{\in \mathbb{R}^{p \times p}} \\ &= \underbrace{\begin{pmatrix} \mathbf{u}_1 & \dots & \mathbf{u}_r \end{pmatrix}}_{\in \mathbb{R}^{n \times r}} \underbrace{\begin{pmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_r & \\ & & & \mathbf{v}_r^T \end{pmatrix}}_{\in \mathbb{R}^{r \times p}} = \sum_{i=1}^r \sigma_i \underbrace{\mathbf{u}_i \times \mathbf{v}_i^T}_{\in \mathbb{R}^{r \times r}} \end{aligned}$$

devient

$$\begin{aligned} \mathbb{A} &= \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \left( \begin{array}{c|c} 3 & 0 \\ 0 & 1 \end{array} \right) \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & \sqrt{2} \end{pmatrix} \\ &\stackrel{r=2}{=} \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \\ &= \frac{3}{2} \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix} + \frac{1}{2} \begin{pmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \end{pmatrix} \end{aligned}$$

Notons que la décomposition n'est pas unique, par exemple avec Octave on trouve

Vecteurs propres unitaires :

$$\mathbb{U} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \quad \mathbb{V} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & \sqrt{2} \end{pmatrix}$$

ce qui donne le même résultat (heureusement !)

```
A=[1 2 0; 2 1 0]
[n,p]=size(A)
r=min(n,p)
```

```
1. AAT=A*A'
[VecAAT,ValAAT]=eig(AAT)
% unsorted list of all eigenvalues
% To produce a sorted vector with the eigenvalues, and re-order the eigenvectors accordingly:
[ee,perm] = sort(diag(abs(ValAAT)), "descend");
ValAAT=diag(ee)
VecAAT=VecAAT(:,perm)
```

```
2. ATA=A'*A
[VecATA,ValATA]=eig(ATA)
% unsorted list of all eigenvalues
% To produce a sorted vector with the eigenvalues, and re-order the eigenvectors accordingly:
[ee,perm] = sort(diag(abs(ValATA)), "descend");
ValATA=diag(ee)
VecATA=VecATA(:,perm)
```

```
3. myS=diag(sqrt(diag(ValATA)),n,p)
myU=VecAAT
myV=VecATA

[UU,SS,VV]=svd(A)
```

```
4. dS=diag(SS)

AA=zeros(5,4);
for i=1:length(dS)
    temp=dS(i)*UU(:,i)*VV(i,:)
    AA+=temp
end
```

### ★ Exercice 6.4 (Tests SVD)

Nous allons appliquer la décomposition SVD à la compression d'images. Nous allons travailler avec des images en niveaux de gris (*grayscale* image en anglais) dont chaque pixel est codé par un entier entre 0 et 255. En conséquence, une image de  $n \times p$  pixels sera représentée par une matrice rectangulaire avec  $n$  lignes et  $p$  colonnes à coefficients dans  $\{0, 1, \dots, 255\}$  contenant des niveaux de gris et réciproquement, toute matrice rectangulaire avec  $n$  lignes et  $p$  colonnes à coefficients dans  $\{0, 1, \dots, 255\}$  peut être visualisée comme une image en niveaux de gris. Voici un exemple :

```
A=ones(100,200);
A(45:55,40:60)=ones(11,21)*255;
```

Octave la transforme en matrice avec la fonction `imread`. On a bien une matrice de taille  $512 \times 512$ . On peut afficher cette matrice comme une image en niveaux de gris comme suit :

```
colormap(gray(256));
imshow(uint8(A));
% uint8(x) convert x to unsigned 8-bit integer type
```

et on obtient

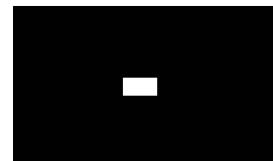


FIGURE 6.1 – Matrice initiale

Considérons la matrice

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

Calculer analytiquement et vérifier numériquement sa décomposition SVD.

Calculer la valeur de  $s$  telle que  $\sigma_i < 10^{-16}$  (le zéro machine) pour  $i = s+1, \dots, r$ . Est-ce plus rentable stocker la matrice  $\mathbb{A}$  ou sa décomposition SVD ?

### Correction

$\mathbb{A} \in \mathbb{R}^{n \times p}$  avec  $n = 3$  et  $p = 2$  donc  $r = 2$ .

Pour calculer la décomposition SVD nous allons calculer les valeurs et vecteurs propres des matrices  $\mathbb{A}\mathbb{A}^T$  et  $\mathbb{A}^T\mathbb{A}$ .

$$\mathbb{A}\mathbb{A}^T = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 2 \end{pmatrix} \quad \mathbb{A}^T\mathbb{A} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

Valeurs propres :

$$\lambda_1 = 2 > \lambda_2 = \lambda_3 = 0$$

$$\lambda_1 = 2 > \lambda_2 = 0$$

Vecteurs propres unitaires :

$$\mathbb{U} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \quad \mathbb{V} = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}$$

Donc

$$\begin{aligned} \mathbb{A} &= \mathbb{U}\mathbb{S}\mathbb{V}^T = \underbrace{\begin{pmatrix} \mathbf{u}_1 & \dots & \mathbf{u}_r & \mathbf{u}_{r+1} & \dots & \mathbf{u}_n \end{pmatrix}}_{\in \mathbb{R}^{n \times n}} \underbrace{\begin{pmatrix} \sigma_1 & & & & & \\ & \ddots & & & & \\ & & \sigma_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix}}_{\in \mathbb{R}^{n \times p}} \underbrace{\begin{pmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_r^T \\ \vdots \\ \mathbf{v}_{r+1}^T \\ \vdots \\ \mathbf{v}_p^T \end{pmatrix}}_{\in \mathbb{R}^{p \times p}} \\ &= \underbrace{\begin{pmatrix} \mathbf{u}_1 & \dots & \mathbf{u}_r \end{pmatrix}}_{\in \mathbb{R}^{n \times r}} \underbrace{\begin{pmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_r & \\ & & & 0 \end{pmatrix}}_{\in \mathbb{R}^{r \times r}} \underbrace{\begin{pmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_r^T \end{pmatrix}}_{\in \mathbb{R}^{r \times p}} = \sum_{i=1}^r \sigma_i \underbrace{\mathbf{u}_i \times \mathbf{v}_i^T}_{\in \mathbb{R}^{r \times r}} \end{aligned}$$

devient

$$\begin{aligned} \mathbb{A} &= \left( \begin{array}{cc|c} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{array} \right) \left( \begin{array}{cc} \sqrt{2} & 0 \\ 0 & 0 \\ \hline 0 & 0 \end{array} \right) \left( \begin{array}{cc} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{array} \right) \\ &\stackrel{r=2}{=} \left( \begin{array}{cc} 0 & 1 \\ 0 & 0 \\ 1 & 0 \end{array} \right) \left( \begin{array}{cc} \sqrt{2} & 0 \\ 0 & 0 \end{array} \right) \left( \begin{array}{cc} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{array} \right) \\ &= \sqrt{2} \left( \begin{array}{cc} 0 & 0 \\ 0 & 0 \\ 1/\sqrt{2} & 1/\sqrt{2} \end{array} \right) + 0 \left( \begin{array}{cc} 1/\sqrt{2} & -1/\sqrt{2} \\ 0 & 0 \\ 0 & 0 \end{array} \right) \stackrel{s=1}{=} \sqrt{2} \left( \begin{array}{cc} 0 & 0 \\ 0 & 0 \\ 1/\sqrt{2} & 1/\sqrt{2} \end{array} \right) \end{aligned}$$

Pour stocker la matrice  $\mathbb{A}$  nous avons besoin de  $n \times p = 3 \times 2 = 6$  valeurs, pour stocker la décomposition SVD nous avons besoin de  $n \times r + r + r \times p = 3 \times 2 + 2 + 2 \times 2 = 12$  valeurs. Cependant, comme  $\sigma_i = 0$  pour  $i = 2$ , nous pouvons reconstruire la matrice  $\mathbb{A}$  en stockant juste une partie de la décomposition SVD et nous avons besoin de  $n \times s + s + s \times p = 3 \times 1 + 1 + 1 \times 2 = 6$  valeurs.

## ★ Exercice 6.5 (Traitement mathématique des images numériques - compression par SVD)

Dans cette exercice nous allons nous intéresser à la manipulation d'images. Nous utiliserons des méthodes basées sur l'algèbre linéaire et l'analyse matricielle.

**Les pixels d'une image :** une image numérique en niveaux de gris (*grayscale* image en anglais) est un tableau de valeurs. Chaque case de ce tableau, qui stocke une valeur, se nomme un pixel. En notant  $n$  le nombre de lignes et  $p$  le nombre de colonnes de l'image, on manipule ainsi un tableau de  $n \times p$  pixels.

La figure ci-dessous montre une visualisation d'un tableau carré avec  $n = p = 512$ , ce qui représente  $512 \times 512 = 2^{18} = 262\,144$  pixels. Les appareils photos numériques peuvent enregistrer des images beaucoup plus grandes, avec plusieurs millions de pixels.

Les valeurs des pixels sont enregistrées dans l'ordinateur ou l'appareil photo numérique sous forme de nombres entiers entre 0 et 255, ce qui fait 256 valeurs possibles pour chaque pixel. La valeur 0 correspond au noir et la valeur 255 correspond au blanc. Les valeurs intermédiaires correspondent à des niveaux de gris allant du noir au blanc.

Pour transformer une image en une matrice il suffit d'indiquer dans notre script :

```
A=imread('lena.jpg');
colormap(gray(256));
A=double(A);
[row,col]=size(A)
```

Octave la transforme en matrice avec la fonction `imread`. On a bien une matrice de taille  $512 \times 512$ . On peut voir Léna<sup>1</sup> avec :

```
colormap(gray(256));
imshow(uint8(A));
% uint8(x) convert x to unsigned 8-bit integer type
```



FIGURE 6.2 – Léna (original)

Tout comme pour la réduction du nombre de pixels, la réduction du nombre de niveaux de gris influe beaucoup sur la qualité de l'image. Afin de réduire au maximum la taille d'une image sans modifier sa qualité, on utilise des méthodes plus complexes de compression d'image.

Soit  $\mathbb{A} \in \mathbb{R}^{n \times p}$  une matrice rectangulaire. Un théorème démontré officiellement en 1936 par C. ECKART et G. YOUNG affirme que toute matrice rectangulaire  $\mathbb{A}$  se décompose sous la forme

$$\mathbb{A} = \mathbb{U} \mathbb{S} \mathbb{V}^T$$

avec  $\mathbb{U} \in \mathbb{R}^{n \times n}$  et  $\mathbb{V} \in \mathbb{R}^{p \times p}$  des matrices orthogonales (i.e.  $\mathbb{U}^{-1} = \mathbb{U}^T$  et  $\mathbb{V}^{-1} = \mathbb{V}^T$ ) et  $\mathbb{S} \in \mathbb{R}^{n \times p}$  une matrice diagonale qui contient les  $r$  valeurs singulières de  $\mathbb{A}$ ,  $r = \min\{n, p\}$ , rangées dans l'ordre décroissant  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r \geq 0$ .

Notons  $\mathbf{u}_i$  et  $\mathbf{v}_i$  les vecteurs colonne des matrices  $\mathbb{U}$  et  $\mathbb{V}$ . La décomposition s'écrit alors

$$\begin{aligned} \mathbb{A} &= \mathbb{U} \mathbb{S} \mathbb{V}^T = \underbrace{\begin{pmatrix} \mathbf{u}_1 & \dots & \mathbf{u}_r & \mathbf{u}_{r+1} & \dots & \mathbf{u}_n \end{pmatrix}}_{\in \mathbb{R}^{n \times n}} \underbrace{\begin{pmatrix} \sigma_1 & & & & & \\ & \ddots & & & & \\ & & \sigma_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix}}_{\in \mathbb{R}^{n \times p}} \underbrace{\begin{pmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_r^T \\ \mathbf{v}_{r+1}^T \\ \vdots \\ \mathbf{v}_p^T \end{pmatrix}}_{\in \mathbb{R}^{p \times p}} \\ &= \underbrace{\begin{pmatrix} \mathbf{u}_1 & \dots & \mathbf{u}_r \end{pmatrix}}_{\in \mathbb{R}^{n \times r}} \underbrace{\begin{pmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_r & \\ & & & \mathbf{v}_r^T \end{pmatrix}}_{\in \mathbb{R}^{r \times p}} = \sum_{i=1}^r \sigma_i \underbrace{\mathbf{u}_i \times \mathbf{v}_i^T}_{\in \mathbb{R}^{r \times r}} \end{aligned}$$

Dans Octave, on peut obtenir la décomposition SVD par la commande

```
[U,S,V]=svd(A);
```

On peut exploiter cette décomposition pour faire des économies de mémoire. En effet, pour stocker la matrice  $\mathbb{A}$  nous avons besoin de  $n \times p$  valeurs, pour stocker la décomposition SVD nous avons besoin de  $n \times r + r + r \times p > n \times p$  valeurs.

1. [http://www.lenna.org/full/l\\_hires.jpg](http://www.lenna.org/full/l_hires.jpg)

Cependant, si nous approchons  $\mathbb{A}$  en ne gardant que les premiers termes de la somme ( sachant que les derniers termes sont multipliés par des  $\sigma_i$  plus petits, voire nuls)

$$\tilde{\mathbb{A}} = \sum_{i=1}^s \sigma_i \underbrace{\mathbf{u}_i \times \mathbf{v}_i^T}_{\in \mathbb{R}^{n \times p}}, \quad \text{où } s < r$$

nous n'avons plus besoin que de  $n \times s + s + s \times p$  valeurs. Cela signifie que pour tout  $s < np/(n+p+1)$  on fait des économies de stockage.

1. Testez la compression avec les différentes valeurs de  $s$  indiquées pour obtenir les images 6.5a-6.5c :
2. La précision d'Octave est de l'ordre de 5 chiffres significatifs. Donc les valeurs singulières significatives doivent avoir un rapport de moins de  $10^{-5}$  avec la valeur maximale  $\sigma_1 = 52517$ , les autres étant considérées comme «erronées». Calculer le nombre de valeurs singulières «significatives», i.e. la plus grande valeur de  $s$  telle que  $\frac{\sigma_i}{\sigma_1} < 10^{-5}$  pour  $i = s+1, \dots, r$  et afficher le résultat comme à la figure 6.5d.



FIGURE 6.3 – SVD

### Correction

```
clear all

% A = imread('lena512.bmp');
% A = imread('lenaTest4.jpg');
A = imread('lena_std.tif');
colormap(gray(256));
A = double(A);
A = (A(:,:,1)+A(:,:,2)+A(:,:,3))/3;
[row,col] = size(A)

figure(1)
colormap(gray(256));
imshow(uint8(A));

[U,S,V]=svd(A);
% [U,S,V]=svd(A,s); % on fixe la troncature

% on fait des économies de stockage si "s" est < a "
    "économie" :
economie = row*col/(row+col+1)

dS=diag(S);
mini=min(dS)
Maxi=max(dS)
dSn=dS/dS(1);
vsSignif=length(dSn(dSn>1.e-5))

for s=[10,100,floor(economie),vsSignif]
    figure(50+s)
    axis image
    X=U(:,1:s)*S(1:s,1:s)*(V(:,1:s))';
    erreur=abs(X-A);
    m=min(min(erreur));
    M=max(max(erreur));
    erreur=255-255/(M-m)*(erreur-m);
    subplot(2,1,1)
    imshow(uint8(X));
    title ( strcat("s=",num2str(s)) );
    %imwrite(uint8(X),strcat("exo6-",num2str(s)," .jpg
        ")," jpg");
    subplot(2,1,2)
    imshow(uint8(erreur));
    title ( strcat("s=",num2str(s)," - erreur dans [
        ",num2str(m)," ; ",num2str(M)," ]) ) ;
end
```

On a  $\max \left\{ s \in [0; r] \mid s < \frac{np}{n+p+1} \right\} = 255$  : on fait des économies de stockage tant qu'on garde au plus les premières 255 valeurs singulières.

La photo de Lena, de taille  $512 \times 512$ , possède 498 valeurs singulières «significatives».

## ★ Exercice 6.6 (Traitement mathématique des images numériques - compression par SVD)

Dans cette exercice nous allons nous intéresser à la manipulation d'images. Nous utiliserons des méthodes basées sur l'algèbre linéaire et l'analyse matricielle.

**Les pixels d'une image :** une image numérique en niveaux de gris (*grayscale* image en anglais) est un tableau de valeurs. Chaque case de ce tableau, qui stocke une valeur, se nomme un pixel. En notant  $n$  le nombre de lignes et  $p$  le nombre de colonnes de l'image, on manipule ainsi un tableau de  $n \times p$  pixels.

La figure ci-dessous montre une visualisation d'un tableau carré avec  $n = p = 512$ , ce qui représente  $512 \times 512 = 2^{18} = 262\,144$  pixels. Les appareils photos numériques peuvent enregistrer des images beaucoup plus grandes, avec plusieurs millions de pixels.

Les valeurs des pixels sont enregistrées dans l'ordinateur ou l'appareil photo numérique sous forme de nombres entiers entre 0 et 255, ce qui fait 256 valeurs possibles pour chaque pixel. La valeur 0 correspond au noir et la valeur 255 correspond au blanc. Les valeurs intermédiaires correspondent à des niveaux de gris allant du noir au blanc.

Pour transformer une image en une matrice il suffit d'indiquer dans notre script :

```
A=imread('flower.png');
colormap(gray(256));
A=double(A);
[row,col]=size(A)
```

Octave la transforme en matrice avec la fonction `imread`. On a bien une matrice de taille  $512 \times 512$ . On peut voir une fleur avec :

```
colormap(gray(256));
imshow(uint8(A));
% uint8(x) convert x to unsigned 8-bit integer type
```



FIGURE 6.4 – Flower (original)

Tout comme pour la réduction du nombre de pixels, la réduction du nombre de niveaux de gris influe beaucoup sur la qualité de l'image. Afin de réduire au maximum la taille d'une image sans modifier sa qualité, on utilise des méthodes plus complexes de compression d'image.

Soit  $\mathbb{A} \in \mathbb{R}^{n \times p}$  une matrice rectangulaire. Un théorème démontré officiellement en 1936 par C. ECKART et G. YOUNG affirme que toute matrice rectangulaire  $\mathbb{A}$  se décompose sous la forme

$$\mathbb{A} = \mathbb{U}\mathbb{S}\mathbb{V}^T$$

avec  $\mathbb{U} \in \mathbb{R}^{n \times n}$  et  $\mathbb{V} \in \mathbb{R}^{p \times p}$  des matrices orthogonales (*i.e.*  $\mathbb{U}^{-1} = \mathbb{U}^T$  et  $\mathbb{V}^{-1} = \mathbb{V}^T$ ) et  $\mathbb{S} \in \mathbb{R}^{n \times p}$  une matrice diagonale qui contient les  $r$  valeurs singulières de  $\mathbb{A}$ ,  $r = \min\{n, p\}$ , rangées dans l'ordre décroissant  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r \geq 0$ .

Notons  $\mathbf{u}_i$  et  $\mathbf{v}_i$  les vecteurs colonne des matrices  $\mathbb{U}$  et  $\mathbb{V}$ . La décomposition s'écrit alors

$$\begin{aligned} \mathbb{A} = \mathbb{U}\mathbb{S}\mathbb{V}^T &= \underbrace{\left(\mathbf{u}_1 \quad \dots \quad \mathbf{u}_r \quad \mathbf{u}_{r+1} \quad \dots \quad \mathbf{u}_n\right)}_{\in \mathbb{R}^{n \times n}} \underbrace{\begin{pmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_r & \\ & & & 0 \end{pmatrix}}_{\in \mathbb{R}^{n \times p}} \underbrace{\left(\mathbf{v}_1^T \quad \vdots \quad \mathbf{v}_r^T \quad \mathbf{v}_{r+1}^T \quad \vdots \quad \mathbf{v}_p^T\right)^T}_{\in \mathbb{R}^{p \times p}} \\ &= \underbrace{\left(\mathbf{u}_1 \quad \dots \quad \mathbf{u}_r\right)}_{\in \mathbb{R}^{n \times r}} \underbrace{\begin{pmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_r & \\ & & & \mathbf{v}_r^T \end{pmatrix}}_{\in \mathbb{R}^{r \times p}} \underbrace{\left(\mathbf{v}_1^T \quad \vdots \quad \mathbf{v}_r^T\right)^T}_{\in \mathbb{R}^{r \times p}} = \sum_{i=1}^r \sigma_i \underbrace{\mathbf{u}_i \times \mathbf{v}_i^T}_{\in \mathbb{R}^{r \times r}} \end{aligned}$$

Dans Octave, on peut obtenir la décomposition SVD par la commande

```
[U,S,V]=svd(A);
```

On peut exploiter cette décomposition pour faire des économies de mémoire. En effet, pour stocker la matrice  $\mathbb{A}$  nous avons besoin de  $n \times p$  valeurs, pour stocker la décomposition SVD nous avons besoin de  $n \times r + r + r \times p > n \times p$  valeurs. Cependant, si nous approchons  $\mathbb{A}$  en ne gardant que les premiers termes de la somme (sachant que les derniers termes sont

multipliés par des  $\sigma_i$  plus petits, voire nuls)

$$\tilde{A} = \sum_{i=1}^s \sigma_i \underbrace{\mathbf{u}_i \times \mathbf{v}_i^T}_{\in \mathbb{R}^{n \times p}}, \quad \text{où } s < r$$

nous n'avons plus besoin que de  $n \times s + s + s \times p$  valeurs. Cela signifie que pour tout  $s < np/(n+p+1)$  on fait des économies de stockage.

1. Testez la compression avec les différentes valeurs de  $s$  indiquées pour obtenir les images 6.5a-6.5c :
2. La précision d'Octave est de l'ordre de 5 chiffres significatifs. Donc les valeurs singulières significatives doivent avoir un rapport de moins de  $10^{-5}$  avec la valeur maximale  $\sigma_1 = 24813$ , les autres étant considérées comme «erronées». Calculer le nombre de valeurs singulières «significatives», i.e. la plus grande valeur de  $s$  telle que  $\frac{\sigma_i}{\sigma_1} < 10^{-5}$  pour  $i = s+1, \dots, r$  et afficher le résultat comme à la figure 6.5d.

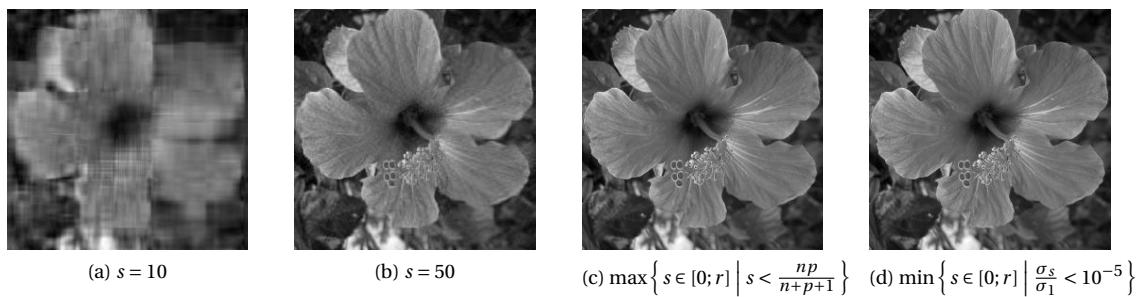


FIGURE 6.5 – SVD

### Correction

```
clear all

A=imread('flower.png');
colormap(gray(256));
A=double(A);
[row,col]=size(A)

figure(1)
colormap(gray(256));
imshow(uint8(A));

[U,S,V]=svd(A);
% [U,S,V]=svd(A,s); % on fixe la troncature

% on fait des économies de stockage si "s" est < a "
    "économie" :
economie=row*col/(row+col+1)

dS=diag(S);
mini=min(dS)
```

```
Maxi=max(dS)
dSn=dS/dS(1);
vsSignif=length(dSn(dSn>1.e-5))

for s=[10,50,floor(economie),vsSignif]
    figure(50+s)
    axis image
    X=U(:,1:s)*S(1:s,1:s)*(V(:,1:s))';
    erreur=abs(X-A);
    m=min(min(erreur));
    M=max(max(erreur));
    erreur=255-255/(M-m)*(erreur-m);
    subplot(2,1,1)
    imshow(uint8(X));
    title ( strcat("s=",num2str(s)) );
    imwrite(uint8(X),strcat("exo6-",num2str(s)," .jpg
        "),'jpg');
    subplot(2,1,2)
    imshow(uint8(erreur));
    title ( strcat("s=",num2str(s)," - erreur dans [
        ",num2str(m)," ; ",num2str(M)," ]) ) ;
end
```

On a  $\max \left\{ s \in [0; r] \mid s < \frac{np}{n+p+1} \right\} = 128$  : on fait des économies de stockage tant qu'on garde au plus les premières 128 valeurs singulières.

La photo de la fleur, de taille  $512 \times 512$ , possède 255 valeurs singulières «significatives».

# Annales

- ★ Contrôle Continu du 5 décembre 2016 : exercices 1.15 à la page 48, 1.54 à la page 79 et 3.21 à la page 148 ;
- ★ Contrôle Terminal 2017 — session 1 : exercices 4.39 à la page 197, 3.18 à la page 144, 1.23 à la page 52 et 6.5 à la page 275 ;
- ★ Contrôle Terminal 2017 — session 2 : exercices 4.39 à la page 197, 3.19 à la page 145, 1.24 à la page 56 et 6.6 à la page 277.
  
- ★ Contrôle Continu du 11 octobre 2017 : bientôt disponible
- ★ Contrôle Continu du 25 octobre 2017 : bientôt disponible
- ★ Contrôle Continu du 18 décembre 2017 : bientôt disponible
- ★ Contrôle Terminal 2018 — session 1 : bientôt disponible
- ★ Contrôle Terminal 2018 — session 2 : bientôt disponible