Diplôme d'ingénieur Cnam

Parcours Informatique et Multimédia

1ère année

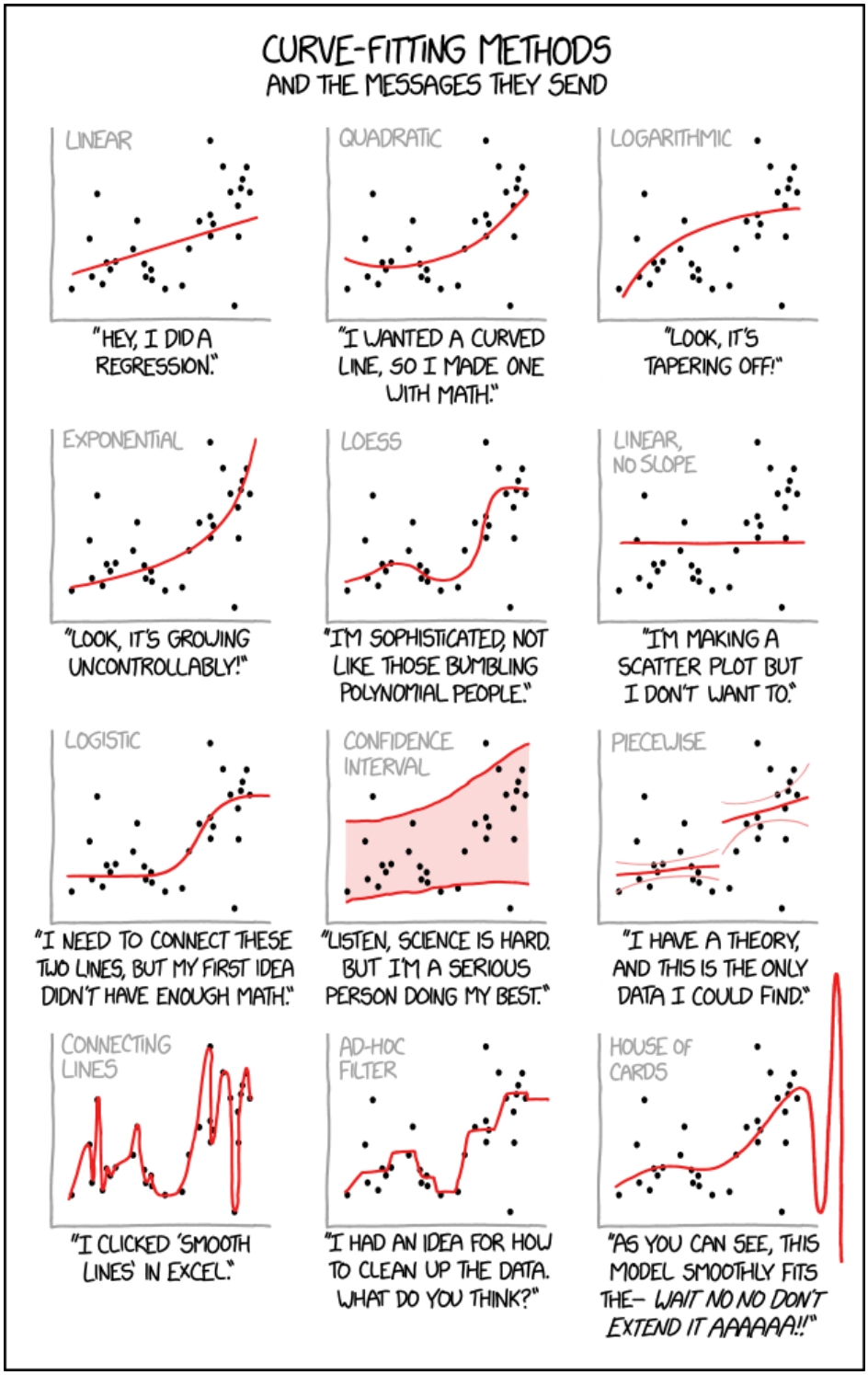
Mathématiques pour les Médias Numériques

Recueil d'exercices corrigés et aide-mémoire.

Gloria Faccanoni

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

Dernière mise-à-jour : Mercredi 31 mai 2023



UNIVERSITÉ DE TOULON LE Cnam Provence-Alpes-Côte d'Azur

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. L'objectif principal de ce cours est la mise en œuvre des connaissances mathématiques acquises les années précédentes dans un contexte de modélisation et initiation à l'analyse numérique pour les médias numériques. 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 détailler le moindre calcul 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.

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. Ici on a cherché, compte tenu des contraintes de volume horaire, des acquis des étudiants 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	70h	20 séances de 3h30

Quelque référence. Il est bon d'utiliser le web via des moteurs de recherche pour trouver des références en format `pdf`. L'encyclopédie *Wikipedia* permet souvent un éclairage intéressant et différent de ce que montrent les manuels universitaires français. Seulement, *internet* ne doit pas être votre seule et unique source d'inspiration. Votre bibliographie doit comprendre d'autres ouvrages scientifiques, comme par exemple :

[1] Howard ANTON, Irl C BIVENS et Stephen DAVIS. *Calculus Early Trascendental*. John Wiley & Sons, 2012.

[2] Frédéric BERTRAND et Myriam MAUMY-BERTRAND. *Initiation à la statistique avec R : Cours, exemples, exercices et problèmes corrigés*. Sciences Sup. Dunod, sept. 2010, 396 pages.

[3] D. CHAPRA C. CHAPRA. *Numerical Methods for Engineers*. McGraw-Hill Education, 2014.

[4] E. Ward CHENEY et David R. KINCAID. *Numerical Mathematics and Computing*. Cengage Learning, 2012.

[5] Bernard GOLDFARB et Catherine PARDOUX. *Introduction à la méthode statistique-6e éd. : Économie, gestion*. Hachette, 2011.

[6] Jaan KIUSALAAS. *Numerical Methods in Engineering with Python 3*. Cambridge University Press, 2013.

[7] Alfio QUARTERONI, Fausto SALERI et Paola GERVASIO. *Calcul scientifique : cours, exercices corrigés et illustrations en MATLAB et Octave*. Springer, 2011.

[8] Jean-Pierre RAMIS, André WARUSFEL, Xavier BUFF, Josselin GARNIER, Emmanuel HALBERSTADT, Thomas LACHAND-ROBERT, François MOULIN et Jacques SAULOY. *Mathématiques Tout-en-un pour la Licence 1-3e éd*. Dunod, 2018.

[9] Timothy SAUER. *Numerical Analysis*. Pearson Education, 2018.

[10] D. STEVEN C. CHAPRA. *Applied Numerical Methods with MATLAB for Engineers and Scientists*. McGraw-Hill Education, 2017.

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
ℹ http://faccanoni.univ-tln.fr

Table des matières

CHAPITRE 1

Background

1.1. Éléments d'analyse matricielle

1.1.1. Généralité

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 $\mathscr{M}_{m,n}(\mathbb{K})$.
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

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

- 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 $\mathscr{M}_n(\mathbb{K})$.
- 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 : $a_{ij} = 0$ pour tout $i = 1, \dots, m$ et tout $j = 1, \dots, n$.
- 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 $\mathrm{Diag}(d_1, d_2, \dots, d_n)$.
- La **MATRICE IDENTITÉ** d'ordre n, notée $\mathbb{I}_n$, est la matrice diagonale $\mathrm{Diag}(\underbrace{1, 1, \dots, 1}_{n \text{ fois}})$.
- 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 $i < j \implies a_{ij} = 0$.
- On appelle matrice **TRANSPOSÉE** de $\mathbb{A}$, notée $\mathbb{A}^T$, la matrice $\mathbb{A} = (a_{ji})_{\substack{1 \leq j \leq n \\ 1 \leq i \leq m}}$. C'est donc une matrice de $\mathscr{M}_{n,m}(\mathbb{R})$ obtenue en échangeant lignes et colonnes de la matrice initiale. Bien évidemment $(\mathbb{A}^T)^T = \mathbb{A}$.
- On appelle matrice **ADJOINTE** (ou CONJUGUÉE TRANSPOSÉE) de $\mathbb{A}$, notée $\mathbb{A}^H$, la matrice $\mathbb{A} = (\overline{a}_{ji})_{\substack{1 \leq j \leq n \\ 1 \leq i \leq m}}$. C'est donc une matrice de $\mathscr{M}_{n,m}(\mathbb{C})$ obtenue en échangeant lignes et colonnes de la matrice initiale et en prenant le nombre complexe conjugué. Bien évidemment $(\mathbb{A}^H)^H = \mathbb{A}$.

- ⋆ Une matrice $\mathbb{A}$ est dite **SYMÉTRIQUE** si $\mathbb{A}^T = \mathbb{A}$, *i.e.* si $a_{ij} = a_{ji}$ pour tout $i \neq j$.
- ⋆ Une matrice $\mathbb{A}$ est dite **HERMITIENNE** ou **AUTOADJOINTE** si $\mathbb{A}^H = \mathbb{A}$, *i.e.* si $\overline{a}_{ij} = a_{ji}$ pour tout $i \neq j$.
- ⋆ 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 : $\mathrm{tr}(\mathbb{A}) \equiv \sum_{i=1}^{n} a_{ii}$. Par conséquent $\mathrm{tr}(\mathbb{A}^T) = \mathrm{tr}(\mathbb{A})$.

On remarque qu'une matrice DIAGONALE est triangulaire supérieure et inférieure (*i.e.* $i \neq j \implies a_{ij} = 0$).

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 $\mathbb{Z}$.
- ⋆ La matrice $\mathbb{U} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 5 & 6 & 7 \\ 0 & 0 & 2 & -1 \\ 0 & 0 & 0 & -5 \end{pmatrix}$ est une matrice triangulaire supérieure.
- ⋆ La matrice $\mathbb{L} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 \\ 5 & -1 & 2 & 0 \\ 7 & 9 & 15 & 4 \end{pmatrix}$ est une matrice triangulaire inférieure.
- ⋆ La matrice $\mathbb{D} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -8 & 0 & 0 \\ 0 & 0 & 7 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$ est une matrice diagonale.
- ⋆ La matrice $\mathbb{I}_4 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$ est la matrice identité d'ordre 4.
- ⋆ La matrice $\mathbb{B} = \begin{pmatrix} 7 & 0 & 8 & 2 \end{pmatrix}$ est une matrice ligne (= vecteur ligne) d'ordre 4.
- ⋆ La matrice $\mathbb{C} = \begin{pmatrix} 7 \\ 0 \\ 9 \end{pmatrix}$ est une matrice colonne (= vecteur colonne) d'ordre 3.
- ⋆ La matrice $\mathbb{A} = \begin{pmatrix} 1 & 5 & -9 \\ 5 & 4 & 0 \\ -9 & 0 & 7 \end{pmatrix}$ est symétrique.
- ⋆ Si $\mathbb{A} = \begin{pmatrix} 1 & -1 & 5 \\ 3 & 0 & 7 \end{pmatrix}$ alors $\mathbb{A}^T = \begin{pmatrix} 1 & 3 \\ -1 & 0 \\ 5 & 7 \end{pmatrix}$.
- ⋆ La trace de la matrice $\mathbb{A} = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 2 & 3 \\ 0 & -1 & -2 \end{pmatrix}$ est $\mathrm{tr}(\mathbb{A}) = a_{11} + a_{22} + a_{33} = 1 + 2 + (-2) = 1$.

1.1.2. Calcul matriciel élémentaire

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 2×3 suivantes :

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

La somme de $\mathbb{A}$ et $\mathbb{B}$ est la matrice 2×3 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.

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

1. $\mathbb{A} + \mathbb{B} = \mathbb{B} + \mathbb{A}$ (commutativité),
2. $\mathbb{A} + (\mathbb{B} + \mathbb{C}) = (\mathbb{A} + \mathbb{B}) + \mathbb{C}$ (associativité),
3. $(\mathbb{A} + \mathbb{B})^T = \mathbb{A}^T + \mathbb{B}^T$
4. $(\mathbb{A} + \mathbb{B})^H = \mathbb{A}^H + \mathbb{B}^H$
5. $\mathrm{tr}(\mathbb{A} + \mathbb{B}) = \mathrm{tr}(\mathbb{A}) + \mathrm{tr}(\mathbb{B})$.

EXEMPLE
Soient les matrices 2×2 suivantes :

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

On a alors

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

De plus,

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

Produit d'une matrice par un scalaire

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

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

Si $\mathbb{A}$ et $\mathbb{B}$ sont deux matrices de même ordre et $\alpha \in \mathbb{K}$ un scalaire, alors $\alpha \cdot (\mathbb{A}+\mathbb{B}) = \alpha \cdot \mathbb{A} + \alpha \cdot \mathbb{B}$ (distributivité).

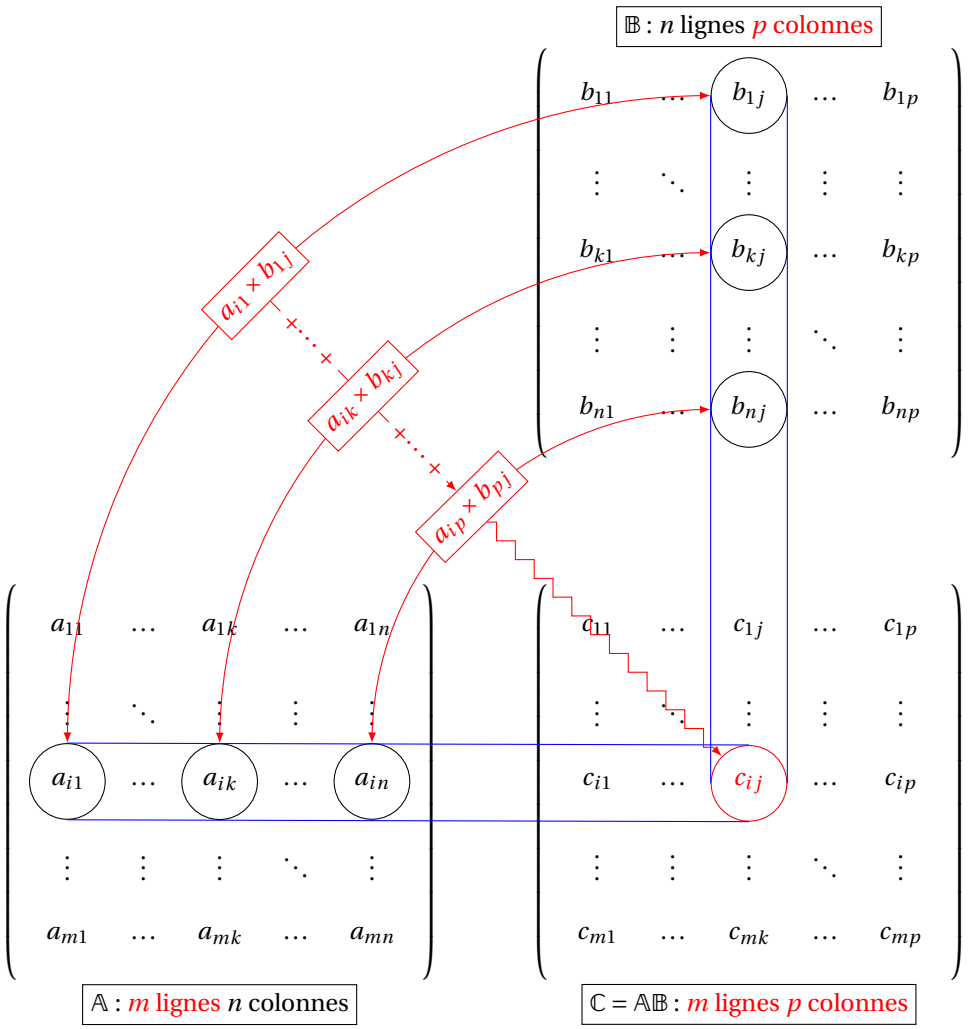
De plus, $(\alpha\mathbb{A})^T = \alpha\mathbb{A}^T$ et $(\alpha\mathbb{A})^H = \overline{\alpha}\mathbb{A}^H$.

EXEMPLE
Si $\alpha = \frac{1}{2}$ et $\mathbb{A} = \begin{pmatrix} 3 & 4 & 2 \\ 1 & 3 & 5 \end{pmatrix}$ alors $\alpha \cdot \mathbb{A} = \begin{pmatrix} 3/2 & 2 & 1 \\ 1/2 & 3/2 & 5/2 \end{pmatrix}$.

Produit de matrices

Si $\mathbb{A} = (a_{ik})_{\substack{1\le i\le m \\ 1\le k\le n}}$ est une matrice $m \times n$ et $\mathbb{B} = (b_{kj})_{\substack{1\le k\le n \\ 1\le j\le p}}$ une matrice $n \times p$, on définit $\mathbb{C} = \mathbb{A}\mathbb{B}$ le **PRODUIT DES MATRICES** $\mathbb{A}$ et $\mathbb{B}$ (dans l'ordre) comme la matrice de dimension $m \times p$ telle que l'élément c_{ij} est le produit scalaire de la ligne i de $\mathbb{A}$ et de la colonne j de $\mathbb{B}$. par

$$c_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj} = a_{i1}b1j + a_{i2}b2j + \cdots + a_{in}bnj$$



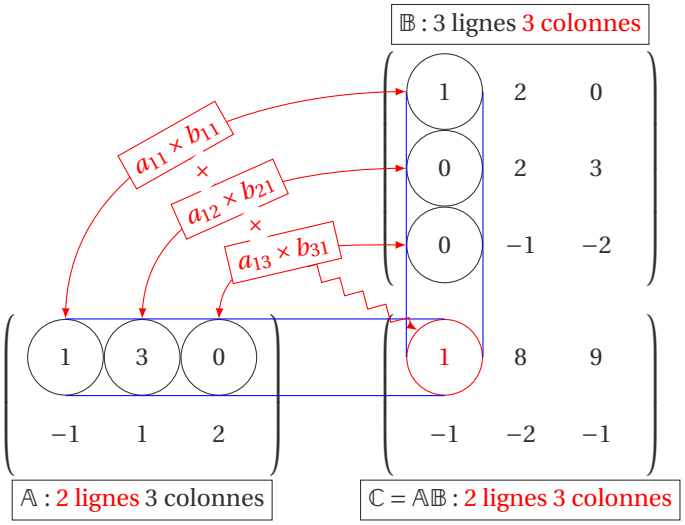
EXEMPLE

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

$$\mathbb{AB} = \begin{pmatrix} 1\times1+3\times0+0\times0 & 1\times2+3\times2+0\times(-1) & 1\times0+3\times3+0\times(-2) \\ -1\times1+1\times0+2\times0 & -1\times2+1\times2+2\times(-1) & -1\times0+1\times3+2\times(-2) \end{pmatrix} = \begin{pmatrix} 1 & 8 & 9 \\ -1 & -2 & -1 \end{pmatrix}.$$



EXEMPLE
Une société commerciale possède deux magasins dont l'aménagement du parc informatique est le suivant :
- ⋆ Magasin 1 : 12 PC, 5 tablettes et 10 smartphones,
- ⋆ Magasin 2 : 17 PC, 6 tablettes et 14 smartphones.

On peut associer à cet équipement la matrice $\mathbb{M} = \left(\begin{smallmatrix} 12 & 5 & 10 \\ 17 & 6 & 14 \end{smallmatrix}\right)$.
La société souhaite améliorer son équipement de la manière suivante :
- ⋆ Magasin 1 : +3 PC, +2 tablettes et +2 smartphones,
- ⋆ Magasin 2 : +5 PC, +3 tablettes et +4 smartphones.

Ce nouvel équipement peut être associé à la matrice $\mathbb{N} = \left(\begin{smallmatrix} 3 & 2 & 2 \\ 5 & 3 & 4 \end{smallmatrix}\right)$.
Le répartition du nouvel aménagement du parc informatique des deux magasins sera donc

$$\mathbb{M}+\mathbb{N} = \begin{pmatrix} 12 & 5 & 10 \\ 17 & 6 & 14 \end{pmatrix} + \begin{pmatrix} 3 & 2 & 2 \\ 5 & 3 & 4 \end{pmatrix} = \begin{pmatrix} 15 & 7 & 12 \\ 22 & 9 & 18 \end{pmatrix}$$

Pour acheter le nouvel équipement, la société commerciale a le choix entre deux fournisseurs :
- ⋆ Fournisseur 1 : 600 e le PC, 180 e la tablette et 60 e le smartphone,
- ⋆ Fournisseur 2 : 550 e le PC, 200 e la tablette et 50 e le smartphone.

On peut associer ces prix à la matrice $\mathbb{P} = \left(\begin{smallmatrix} 600 & 550 \\ 180 & 200 \\ 60 & 50 \end{smallmatrix}\right)$.
On obtient les prix du nouvel aménagement selon les magasins et selon les fournisseurs en calculant

$$\mathbb{NP} = \begin{pmatrix} 3 & 2 & 2 \\ 5 & 3 & 4 \end{pmatrix} \begin{pmatrix} 600 & 550 \\ 180 & 200 \\ 60 & 50 \end{pmatrix} = \begin{pmatrix} 3\times 600+2\times 180+2\times 60 & 3\times 550+2\times 200+2\times 50 \\ 5\times 600+3\times 180+4\times 60 & 5\times 550+3\times 200+4\times 50 \end{pmatrix} = \begin{pmatrix} 2280 & 2150 \\ 3780 & 3550 \end{pmatrix}$$

Par exemple, le prix de l'investissement pour le magasin 1 est de 2280 e avec le fournisseur 1 et de 2150 e avec le fournisseur 2.

ATTENTION
$\mathbb{AB} \neq \mathbb{BA}$ 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 $\mathbb{AB}$ est défini, c'est une matrice d'ordre $m \times n$.
Qu'en est-il du produit $\mathbb{BA}$? Il faut distinguer trois cas :
- ⋆ si $m \neq n$ le produit $\mathbb{BA}$ n'est pas défini ;
- ⋆ si $m = n$ mais $p \neq n$, le produit $\mathbb{AB}$ est défini et c'est une matrice d'ordre $m \times n$ tandis que le produit $\mathbb{BA}$ est défini mais c'est une matrice d'ordre $p \times p$ donc $\mathbb{AB} \neq \mathbb{BA}$;
- ⋆ si $m = n = p$, $\mathbb{A}$ et $\mathbb{B}$ sont deux matrices carrées d'ordre m. Les produits $\mathbb{AB}$ et $\mathbb{BA}$ sont aussi carrés et d'ordre m mais là encore, en général, $\mathbb{AB} \neq \mathbb{BA}$;

EXEMPLE
Soient les matrices

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

On obtient

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

Si les dimensions sont compatibles, on a les propriétés suivantes :

1. $\mathbb{A}(\mathbb{BC}) = (\mathbb{AB})\mathbb{C}$ (associativité)
2. $\mathbb{A}(\mathbb{B}+\mathbb{C}) = \mathbb{AB}+\mathbb{AC}$ (distributivité)
3. $\mathbb{A}\mathbb{I}_n = \mathbb{I}_n\mathbb{A} = \mathbb{A}$
4. $(\mathbb{AB})^T = \mathbb{B}^T\mathbb{A}^T$
5. $(\mathbb{AB})^H = \mathbb{B}^H\mathbb{A}^H$
6. $\mathrm{tr}(\mathbb{AB}) = \mathrm{tr}(\mathbb{BA})$

Puissance d'une matrice

Si $\mathbb{A}$ est une matrice carrée, on note $\mathbb{B} = \mathbb{A}^q$ (pour $q \geq 2$) la matrice définie par

$$\mathbb{B} \equiv \underbrace{\mathbb{A} \times \mathbb{A} \times \cdots \times \mathbb{A}}_{\text{q fois}}.$$

Il s'agit du produit matriciel de $\mathbb{A}$ par elle même q fois par conséquent, en générale, $b_{ij} \neq (a_{ij})^q$.
Si la matrice est diagonale, *i.e.* si $a_{ij} = 0$ pour $i \neq j$, alors $b_{ij} = (a_{ij})^q$.

Inverse d'une matrice carrée

Une matrice carrée $\mathbb{A} \in \mathscr{M}_n(\mathbb{K})$ est dite **INVERSIBLE** (ou régulière) s'il existe une matrice $\mathbb{B} \in \mathscr{M}_n(\mathbb{K})$ telle que

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

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

- ⋆ Si une matrice est non inversible (*i.e.* il n'existe pas $\mathbb{A}^{-1}$), on dit qu'elle est SINGULIÈRE.
- ⋆ Une matrice carrée $\mathbb{A}$ est dite ORTHOGONALE si elle est inversible et $\mathbb{A}^T\mathbb{A} = \mathbb{A}\mathbb{A}^T = \mathbb{I}_n$, *i.e.* si $\mathbb{A}^T = \mathbb{A}^{-1}$.
- ⋆ Une matrice carrée $\mathbb{A}$ 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 $\mathbb{A}^H = \mathbb{A}^{-1}$.

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

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

EXEMPLE
Considérons les matrices

$$\mathbb{A} = \begin{pmatrix} 1 & 1 \\ 2 & 4 \end{pmatrix} \qquad \mathbb{B} = \begin{pmatrix} 2 & -\frac{1}{2} \\ -1 & \frac{1}{2} \end{pmatrix}.$$

On a

$$\mathbb{A}\mathbb{B} = \begin{pmatrix} 1 & 1 \\ 2 & 4 \end{pmatrix}\begin{pmatrix} 2 & -\frac{1}{2} \\ -1 & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \mathbb{I}_2 \qquad \mathbb{B}\mathbb{A} = \begin{pmatrix} 2 & -\frac{1}{2} \\ -1 & \frac{1}{2} \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 2 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \mathbb{I}_2$$

On dit que $\mathbb{B}$ est la matrice inverse de $\mathbb{A}$ et réciproquement.

Remarque
Matrice inverse et systèmes linéaires Il est fréquent, dans toutes les disciplines scientifiques, de devoir résoudre des systèmes linéaires.
Tout système linéaire de n équations à n inconnues peut s'écrire sous la forme matricielle $\mathbb{A}\mathbf{x} = \mathbf{b}$; $\mathbb{A}$ est une matrice carrée de dimension n 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.
Si $\mathbb{A}$ est inversible alors ce système possède une unique solution $\mathbf{x}$ donnée par $\mathbf{x} = \mathbb{A}^{-1}\mathbf{b}$ car

$$\mathbb{A}\mathbf{x} = \mathbf{b} \iff \mathbb{A}^{-1}\mathbb{A}\mathbf{x} = \mathbb{A}^{-1}\mathbf{b} \iff \mathbf{x} = \mathbb{A}^{-1}\mathbf{b}.$$

EXEMPLE
Considérons le système linéaire

$$\begin{cases} 2x + 3y = 15 \\ 3x + 4y = 12 \end{cases}$$

Si on pose $\mathbb{A} = \left(\begin{smallmatrix} 2 & 3 \\ 3 & 4 \end{smallmatrix}\right)$, $\mathbf{x} = \left(\begin{smallmatrix} x \\ y \end{smallmatrix}\right)$ et $\mathbf{b} = \left(\begin{smallmatrix} 15 \\ 12 \end{smallmatrix}\right)$, alors le produit matriciel $\mathbb{A}\mathbf{x}$ donne le vecteur colonne

$$\mathbb{A}\mathbf{x} = \begin{pmatrix} 2 & 3 \\ 3 & 4 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 2x + 3y \\ 3x + 4y \end{pmatrix}$$

ainsi le système peut s'écrire sous forme matricielle $\mathbb{A}\mathbf{x} = \mathbf{b}$.
On cherche donc à calculer $\mathbb{A}^{-1}$, *i.e.* on cherche a, b, c, d tels que

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} 2 & 3 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 2 & 3 \\ 3 & 4 \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

On a

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} 2 & 3 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 2a + 3b & 3a + 4b \\ 2c + 3d & 3c + 4d \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$
$$\begin{pmatrix} 2 & 3 \\ 3 & 4 \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 2a + 3c & 2b + 3d \\ 3a + 4c & 3b + 4d \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

si et seulement si $a=-4$, $b=c=3$ et $d=-2$ ainsi

$$\mathbf{x}=\mathbb{A}^{-1}\mathbf{b}=\begin{pmatrix}-4 & 3\\ 3 & -2\end{pmatrix}\begin{pmatrix}15\\ 12\end{pmatrix}=\begin{pmatrix}-24\\ 21\end{pmatrix}$$

Cet exemple montre le lien entre résolution d'un système linéaire et calcul d'une matrice inverse. Cependant, pour calculer la solution du système initiale de 2 équations à 2 inconnues, on doit calculer les 4 coefficients de $\mathbb{A}^{-1}$ et pour cela on doit résoudre un système linéaire de 8 équations et 4 inconnues... ce n'est pas la bonne stratégie!

1.1.3. Définition et calcul pratique d'un déterminant

Le déterminant est un nombre que l'on associe à n vecteurs de $\mathbb{R}^n$. Il correspond au volume du parallélépipède engendré par ces n vecteurs. On peut aussi définir le déterminant d'une matrice $\mathbb{A}$ comme le déterminant des vecteurs qui composent ses colonnes. Le déterminant permet de savoir si une matrice est inversible ou pas, et de façon plus générale, joue un rôle important dans le calcul matriciel et la résolution de systèmes linéaires.

Définition 1.1 (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, $1\le i,j\le n$, 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
$$\det(\mathbb{A})\equiv\sum_{j=1}^{n}(-1)^{i+j}a_{ij}\det(\mathbb{A}_{ij}) \qquad \text{quelque soit la ligne } i \text{ fixée, } 1\le i\le n,$$
ou, de manière équivalente, le nombre
$$\det(\mathbb{A})\equiv\sum_{i=1}^{n}(-1)^{i+j}a_{ij}\det(\mathbb{A}_{ij}) \qquad \text{quelque soit la colonne } j \text{ fixée, } 1\le j\le n.$$

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 (Déterminant d'une matrice d'ordre 2 — méthode de Laplace)
Soit la matrice

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

alors

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

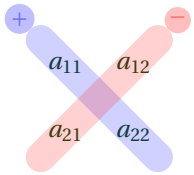
On peut calculer $\det(\mathbb{A})$ par l'une des formules suivantes :

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

Astuce (Déterminant d'une matrice d'ordre 2 — méthode pratique)
Soit $\mathbb{A}$ une matrice carrée d'ordre $n = 2$. Sans appliquer la méthode de Laplace, nous pouvons nous rappeler du déterminant par le schéma suivant :

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



EXEMPLE

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

EXEMPLE (DÉTERMINANT D'UNE MATRICE D'ORDRE 3 — MÉTHODE DE LAPLACE)
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

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

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

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

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

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

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.

EXEMPLE
Soit la matrice

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

alors

$$\det(\mathbb{A}_{11}) = \det\begin{pmatrix} 2 & 0 \\ 3 & 5 \end{pmatrix} = 10, \qquad \det(\mathbb{A}_{12}) = \det\begin{pmatrix} 0 & 0 \\ 0 & 5 \end{pmatrix} = 0, \qquad \det(\mathbb{A}_{13}) = \det\begin{pmatrix} 0 & 2 \\ 0 & 3 \end{pmatrix} = 0,$$

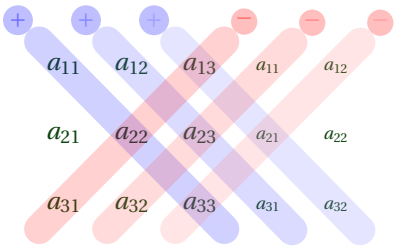
$$\det(\mathbb{A}_{21}) = \det\begin{pmatrix} 0 & 1 \\ 3 & 5 \end{pmatrix} = -3, \qquad \det(\mathbb{A}_{22}) = \det\begin{pmatrix} 1 & 1 \\ 0 & 5 \end{pmatrix} = 5, \qquad \det(\mathbb{A}_{23}) = \det\begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix} = 3,$$
$$\det(\mathbb{A}_{31}) = \det\begin{pmatrix} 0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

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

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

Astuce (Déterminant d'une matrice d'ordre 3 — méthode pratique (règle de SARRUS))
Soit $\mathbb{A}$ une matrice carrée d'ordre $n = 3$. Sans appliquer la méthode de Laplace, nous pouvons nous rappeler du déterminant par le schéma suivant :

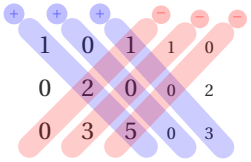
$$\det(\mathbb{A}) = \det\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \left(a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32}\right) - \left(a_{13}a_{22}a_{31} + a_{11}a_{23}a_{32} + a_{12}a_{21}a_{33}\right)$$



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



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

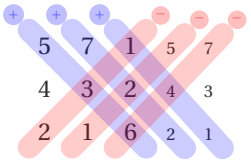
Si on utilise la définition (règle de LAPLACE), en développant selon la première colonne on obtient

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

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

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

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

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

On a les propriétés suivantes :

1. $\mathbb{A}$ est inversible si et seulement si $\det(\mathbb{A}) \neq 0$,
2. $\det(\mathbb{A}^{-1}) = \dfrac{1}{\det(\mathbb{A})}$,
3. $\det(\mathbb{A}^T) = \det(\mathbb{A})$,
4. $\det(\mathbb{A}\mathbb{B}) = \det(\mathbb{A}) \cdot \det(\mathbb{B})$
5. le déterminant d'une matrice triangulaire est égal au produit des éléments diagonaux,
6. 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 α ;
- ⋆ 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.*

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

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

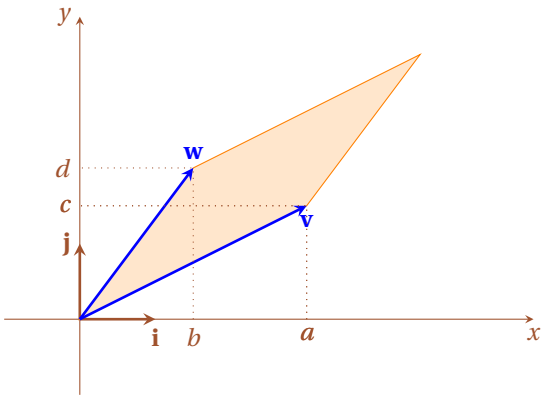
EXEMPLE
Soit la matrice

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

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

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

En dimension 2 les déterminants correspondent à des aires et en dimension 3 à des volumes.
Considérons deux vecteurs $\mathbf{v} = \binom{a}{c}$ et $\mathbf{w} = \binom{b}{d}$ du plan $\mathbb{R}^2$. Ces deux vecteurs déterminent un parallélogramme :



L'aire du parallélogramme est donnée par la valeur absolue du déterminant de la matrice dont les colonnes sont **v** et **w** :

$$\text{Aire} = \left|\det\begin{pmatrix}a & c\\ b & d\end{pmatrix}\right|$$

Théorème 1.2
$\mathbb{A}$ est inversible si et seulement si $\det(\mathbb{A}) \neq 0$.

Propriété 1.3
- $\det(\mathbb{A}^T) = \det(\mathbb{A})$,
- $\det(\mathbb{A}^H) = \overline{\det(\mathbb{A})}$,
- $\det(\mathbb{A}^{-1}) = \dfrac{1}{\det(\mathbb{A})}$,
- $\det(\mathbb{A}\mathbb{B}) = \det(\mathbb{A}) \cdot \det(\mathbb{B})$.

Définition 1.4 (Rang)
Le RANG d'une matrice quelconque $\mathbb{A} \in \mathscr{M}_{m,n}$, 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 $\mathbb{A} \in \mathscr{M}_{m,n}$. 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 2×3 donc $s \leq \min\{2, 3\}$ soit encore $s = 0$, 1 ou 2 ;

- il existe au moins un élément de $\mathbb{A}$ différent de zéro, donc $s \neq 0$ soit encore $s = 1$ ou 2 ; pour qu'il soit 2 il faut trouver une sous-matrice de dimension 2 inversible :
  - comme le déterminant de la sous-matrice composée de la première et de la deuxième colonne est nul, on ne peut pas conclure car je peux encore trouver une autre sous-matrice de dimension 2 inversible ;
  - 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 3×3 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 (car $L_1 = -L_3$) 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 1.5 (Opérations élémentaires sur les lignes d'une matrices)
Les opérations (ou manipulations) élémentaires sur les lignes d'une matrices $\mathbb{M} \in \mathscr{M}_{m,n}$ sont

- la multiplication d'une ligne L_i par un scalaire non nul α :

$$L_i \leftarrow \alpha L_i;$$

- l'addition d'un multiple d'une ligne αL_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 $\mathbb{M} \in \mathscr{M}_{m,n}$ 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 $\mathbb{M} \in \mathscr{M}_{4,3}$ 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 1.6 (Opérations élémentaires sur les colonnes d'une matrices)
Les opérations élémentaires sur les colonnes d'une matrice $\mathbb{M} \in \mathscr{M}_{m,n}$ sont

- la multiplication d'une colonne C_i par un scalaire α non nul :

$$C_i \leftarrow \alpha C_i;$$

- l'addition d'un multiple d'une colonne αC_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 1.7 (Matrices ÉQUIVALENTES)
Deux matrices sont dites ÉQUIVALENTES si on peut passer de l'une à l'autre par des opérations élémentaires.

Théorème 1.8
Deux matrices équivalentes ont le même rang.

1.1.4. 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 1.9 (p-norme ou norme de HÖLDER)
On définit la p-norme (ou norme de HÖLDER) par

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

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 1.10 (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 $\mathbf{x}$.

1.2. Espaces vectoriels

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

Définition 1.11 (Espace vectoriel)
Un ESPACE VECTORIELsur 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

1. 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 $\alpha \cdot (x_1, x_2, \dots, x_n) = (\alpha x_1, \alpha x_2, \dots, \alpha x_n)$.
2. L'ensemble $\mathbb{R}_n[x] = \left\{ p(x) = \sum_{i=1}^{n+1} \alpha_i x^{i-1} \;\middle|\; \alpha_i \in \mathbb{R} \text{ ou } \mathbb{C} \right\}$ 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 1.12 (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 1.13 (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 $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_p$.

EXEMPLE

Considérons les trois vecteurs

$$\mathbf{u}_1 = \begin{pmatrix} -1 \\ -2 \\ -3 \end{pmatrix}, \qquad \mathbf{u}_2 = \begin{pmatrix} 0 \\ 2 \\ -1 \end{pmatrix}, \qquad \mathbf{u}_3 = \begin{pmatrix} -1 \\ 0 \\ -4 \end{pmatrix}.$$

Montrons que $\mathbf{u}_3$ est combinaison linéaire des vecteurs $\mathbf{u}_1$ et $\mathbf{u}_2$.
Pour prouver qu'un vecteur $\mathbf{v}$ est une combinaison linéaire des vecteurs $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_p$ il faut montrer qu'il existe p constantes $\alpha_1, \alpha_2, \dots, \alpha_p$ telles que

$$\mathbf{v} = \alpha_1 \mathbf{u}_1 + \alpha_2 \mathbf{u}_2 + \cdots + \alpha_p \mathbf{u}_p.$$

On cherche alors a et b réels tels que

$$\mathbf{u}_3 = a\mathbf{u}_1 + b\mathbf{u}_2,$$

ce qui donne

$$\begin{cases} -1 = -a, \\ 0 = -2a + 2b, \\ -4 = -3a - b, \end{cases} \iff a = b = 1.$$

Par conséquent $\mathbf{u}_3$ est combinaison linéaire des vecteurs $\mathbf{u}_1$ et $\mathbf{u}_2$ car $\mathbf{u}_3 = \mathbf{u}_1 + \mathbf{u}_2$.

EXEMPLE

Considérons les trois polynômes

$$q_1(x) = 1 + x, \qquad q_2(x) = x + x^2, \qquad q_3(x) = 1 - x^2.$$

Montrons que q_3 est combinaison linéaire des polynômes q_1 et q_2.

Pour prouver qu'un polynôme v est une combinaison linéaire des polynômes $q_1, q_2, \dots, q_p$ il faut montrer qu'il existe p constantes $\alpha_1, \alpha_2, \dots, \alpha_p$ telles que

$$v(x) = \alpha_1 q_1(x) + \alpha_2 q_2(x) + \cdots + \alpha_p q_p(x) \qquad \forall x \in \mathbb{R}.$$

On cherche alors a et b réels tels que

$$q_3(x) = a q_1(x) + b q_2(x), \qquad \forall x \in \mathbb{R}$$

ce qui donne

$$1 - x^2 = a(1+x) + b(x+x^2) \qquad \forall x \in \mathbb{R}$$

soit encore

$$(a-1) + (a+b)x + (1+b)x^2 = 0 \qquad \forall x \in \mathbb{R}.$$

On cherche a et b réels tels que

$$\begin{cases} a - 1 = 0, \\ a + b = 0, \\ 1 + b = 0, \end{cases} \iff a = -b = 1.$$

Par conséquent q_3 est combinaison linéaire des polynômes q_1 et q_2 car $q_3(x) = q_1(x) - q_2(x)$ pour tout $x \in \mathbb{R}$.

Définition 1.14 (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 $\mathrm{Vect}\{\mathbf{u}_1, \dots, \mathbf{u}_p\}$:

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

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

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

EXEMPLE
- $\mathrm{Vect}\{\mathbf{0}_E\} = \{\mathbf{0}_E\}$
- $\mathrm{Vect}\left\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \right\} = \left\{ a \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + b \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \;\middle|\; a, b \in \mathbb{R} \right\} = \left\{ \begin{pmatrix} a & b \\ b & a \end{pmatrix} \;\middle|\; a, b \in \mathbb{R} \right\}.$

Définition 1.15 (Famille libre, famille génératrice, base)
Soit $p \in \mathbb{N}^*$, E un espace vectoriel et $\mathscr{F} = \{\mathbf{u}_1, \dots, \mathbf{u}_p\}$ une famille de vecteurs de E. On dit que la famille $\mathscr{F}$ 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

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

Dans le cas contraire la famille est dite liée.

Pour montrer qu'une famille de plus de deux vecteurs est libre, on sera amené à résoudre le système linéaire correspondant, qui est un système homogène : la famille est libre si et seulement si le système admet uniquement la solution nulle.

… 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 $\mathscr{F}$, on écrit $\mathrm{coord}(\mathbf{u}, \mathscr{F}) = (\alpha_1, \dots, \alpha_p)$ 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.

Attention à ne pas confondre DIMENSION *et* CARDINAL *: dans un espace vectoriel de dimension n, toutes les bases ont le même cardinal (*i.e. *même nombre d'éléments), mais il ne faut pas parler de cardinal d'un espace vectoriel, ni de dimension d'une base.*

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.
- La famille $\{\mathbf{u}=(1,0,-1),\mathbf{v}=(2,3,5),\mathbf{w}=(-1,0,1)\}$ de vecteurs de $\mathbb{R}^3$ n'est pas libre car $\mathbf{w}=-\mathbf{u}$.
- La famille $\{\mathbf{u}=(1,1,-1),\mathbf{v}=(2,-1,2),\mathbf{w}=(3,0,1)\}$ de vecteurs de $\mathbb{R}^3$ n'est pas libre car $\mathbf{w}=\mathbf{u}+\mathbf{v}$.

Théorème 1.16
Dans un espace vectoriel E de dimension n, une FAMILLE GÉNÉRATRICE a au moins n éléments.
Si elle a plus de n éléments, alors elle n'est pas libre mais on peut en extraire une sous-famille libre de cardinal n qui est alors une base de E.
Si elle a exactement n éléments, c'est une base de E.

Théorème 1.17 (de la base incomplète)
Dans un espace vectoriel E de dimension n, une FAMILLE LIBRE a au plus n éléments.
Si elle a moins de n éléments, alors elle n'est pas une base de E mais on peut la compléter de façon à obtenir une base.
Si elle a exactement n éléments, c'est une base de E.

Théorème 1.18 (de la dimension)
Soit $\mathscr{F}$ une famille d'éléments de E de dimension finie n. Les propriétés suivantes sont équivalentes :
1. $\mathscr{F}$ est une base de E
2. $\mathscr{F}$ est libre et contient n éléments
3. $\mathscr{F}$ est génératrice de E et contient n éléments
4. $\mathscr{F}$ est libre et génératrice de E

ATTENTION
On utilise ce théorème principalement pour montrer qu'une famille $\mathscr{F}$ est une base de E. On utilisera surtout les implications suivantes (avec E de dimension n) :
- si $\mathscr{F}$ est libre et de cardinal n alors $\mathscr{F}$ est une base de E
- si $\mathscr{F}$ est libre et génératrice de E alors $\mathscr{F}$ est une base 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 $\mathscr{B}=\{(1,0,\dots,0);(0,1,\dots,0);\dots;(0,0,\dots,1)\}$ est une base, appelée BASE CANONIQUE de $\mathbb{R}^n$, car pour tout vecteur $\mathbf{u}\in\mathbb{R}^n$, $\mathbf{u}=(u_1,u_2,\dots,u_n)=u_1\cdot(1,0,\dots,0)+u_2\cdot(0,1,\dots,0)+\cdots+u_n\cdot(0,0,\dots,1)$ 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 $\mathscr{C}=\{1,x,x^2,\dots,x^n\}$ est appelée BASE CANONIQUE de $\mathbb{R}_n[x]$ car, pour tout polynôme $p\in\mathbb{R}_n[x]$, $p(x)=a_0+a_1x+a_2x^2+\cdots+a_nx^n$ 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 $\mathscr{B}$ de E (qu'on peut noter $\mathrm{coord}(\mathbf{u},\mathscr{B})$).

EXEMPLE
Le polynôme $p(x)=a+bx+cx^2$ a pour coordonnées (a,b,c) dans la base canonique $\mathscr{C}=\{1,x,x^2\}$ 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.

Astuce
Si $\mathscr{F} = \{\mathbf{e}_1, \dots, \mathbf{e}_p\}$ est une famille génératrice d'un espace vectoriel E et si un des vecteurs de $\mathscr{F}$ (par exemple $\mathbf{e}_1$) est une combinaison linéaire des autres vecteurs de $\mathscr{F}$, alors $\mathscr{F} \setminus \{\mathbf{e}_1\}$ est encore une famille génératrice de E. Ce résultat permet en particulier de construire une base d'un espace vectoriel connaissant une famille génératrice de cet espace.

Astuce
Pour déterminer le rang d'une famille de vecteurs $\mathscr{F} = \{\mathbf{e}_1, \dots, \mathbf{e}_p\}$ d'un espace vectoriel E on cherche d'éventuelles relations entre les vecteurs $\mathbf{e}_1, \dots, \mathbf{e}_p$:

- si la famille est libre, on en déduit que $\text{rg}(\mathscr{F}) = p$
- sinon, on cherche à exprimer un vecteur $\mathbf{e}_i$ comme combinaison linéaire des autres vecteurs et on «élimine» ce vecteur de la famille ; on procède ainsi jusqu'à obtenir une famille libre contenue dans $\mathscr{F}$.

Avant de commencer une recherche précise, on peut encadrer $\text{rg}(\mathscr{F})$. Ainsi

- $\text{rg}(\mathscr{F}) \leq \min\{p, \dim(E)\}$ si E est de dimension finie ;
- si $\mathscr{F}$ contient au moins deux vecteurs non colinéaires alors $\text{rg}(\mathscr{F}) \geq 2$;
- si $\mathscr{F}$ contient une famille libre de q vecteurs, alors $\text{rg}(\mathscr{F}) \geq q$.

Définition 1.19 (Rang d'une famille de vecteurs)
Soit E un espace vectoriel et $\mathscr{F} = \{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$ une famille d'éléments de E. On appelle RANG DE $\mathscr{F}$, et on note $\text{rg}(\mathscr{F})$, la dimension du sous-espace vectoriel de E engendré par $\mathscr{F}$:

$$\text{rg}(\mathscr{F}) = \dim(\text{Vect}\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}).$$

ATTENTION
Le rang d'une matrice $\mathbb{A}$ est le rang des vecteurs colonnes de $\mathbb{A}$, c'est-à-dire la dimension du sous-espace vectoriel qu'ils engendrent. Donc

$$\text{rg}(\mathscr{F}) = \text{rg}([\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n]),$$

où $[\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n]$ est la matrice dont les colonnes sont les vecteurs de la famille $\mathscr{F}$.

1.3. Systèmes linéaires et calcul pratique de la matrice inverse

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

$$(S) \quad \begin{cases} a_{11}x_1 & + & \dots & + & a_{1p}x_p & = & b_1, \\ \vdots & & & & \vdots & & \vdots \\ a_{n1}x_1 & + & \dots & + & a_{np}x_p & = & b_n. \end{cases}$$

- 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.
- Deux systèmes sont ÉQUIVALENTS s'ils admettent les mêmes solutions.
- Le SYSTÈME HOMOGÈNE associé à (S) est le système obtenu en remplaçant les b_i par 0.
- Un système est CARRÉ si $n = p$.

Si on note

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \qquad \mathbb{A} = \begin{pmatrix} a_{11} & \dots & a_{1p} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{np} \end{pmatrix}$$

le système (S) est équivalent à l'écriture matricielle $\mathbb{A}\mathbf{x} = \mathbf{b}$.
Si on ajoute le vecteur-colonne des seconds membres $\mathbf{b}$ à la matrice des coefficients $\mathbb{A}$, on obtient ce qu'on appelle la matrice augmentée que l'on note $[\mathbb{A}|\mathbf{b}]$.

Système échelonné (ou triangulaire supérieur)

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 :

$$\begin{cases} 5x_1 - x_2 - x_3 + 2x_4 + x_6 = b_1 \\ 3x_3 - x_4 + 2x_5 = b_2 \\ -x_5 + x_6 = b_3 \\ 5x_6 = b_4 \\ 0 = b_5 \end{cases}$$

La résolution d'un système linéaire $\mathbb{A}\mathbf{x} = \mathbf{b}$ échelonné est simple car, la matrice lui associée étant triangulaire supérieure, on utilise la relation de récurrence (dite *par remontée*)

$$\begin{cases} x_n = \dfrac{b_n}{a_{nn}}, \\ x_i = \dfrac{1}{a_{ii}}\left(b_i - \displaystyle\sum_{j=i+1}^{n} a_{ij}x_j\right), \quad \text{pour } i = n-1, n-2 \ldots, 1 \end{cases}$$

EXEMPLE
Résolution du système triangulaire supérieur : $\begin{cases} x_1+x_2+x_3=6, \\ x_2+x_3=5, \\ x_3=3. \end{cases}$

$$\begin{cases} x_3 = \frac{b_3}{a_{33}} = \frac{3}{1}, \\ x_2 = x_i = \frac{1}{a_{22}}(b_2 - a_{23}x_3) = \frac{1}{1}(5 - x_3) = 2 \\ x_1 = x_i = \frac{1}{a_{11}}(b_1 - a_{12}x_2 - a_{13}x_3) = \frac{1}{1}(6 - x_2 - x_3) = 1. \end{cases}$$

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.

Par conséquent, un système échelonné permet d'établir si le système est possible ou impossible comme dans l'exemple suivant.

EXEMPLE
Établir si les trois systèmes linéaires suivantes sont impossibles ou possibles et, dans ce cas, calculer la/les solution(s).

$$(1)\begin{cases} x+y+z=6, \\ y+z=5, \\ z=3. \end{cases} \qquad (2)\begin{cases} x+y+z=6, \\ y+z=5, \\ 0=0. \end{cases} \qquad (3)\begin{cases} x+y+z=6, \\ y+z=5, \\ 0=3. \end{cases}$$

(1) Ce système est possible et admet une et une seule solution : en partant de la dernière ligne et en remontant, on obtient

$$\begin{aligned} z &= 3, \\ y &= 5 - z = 5 - 3 = 2, \\ x &= 6 - y - z = 6 - 2 - 3 = 1. \end{aligned}$$

(2) Ce système est possible et admet une infinité de solutions : en partant de la dernière ligne et en remontant, on obtient

$$\begin{aligned} z &= \kappa \in \mathbb{R}, \\ y &= 5 - \kappa, \\ x &= 6 - y - z = 1. \end{aligned}$$

(3) Le système n'a pas de solution car aucune valeur de z permet de résoudre $0z = 3$.

Systèmes équivalents et opérations élémentaires

Deux systèmes sont ÉQUIVALENTS s'ils ont les mêmes solutions. Les opérations suivantes donnent des systèmes équivalents :

- ⋆ remplacer une ligne par elle même ± un multiple d'une autre ligne

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

comme par exemple

$$\begin{cases} x+y+ \ \ z=6, \\ \ \ y+ \ \ z=5, \\ \ \ y+2z=8, \end{cases} \xrightarrow{L_3 \leftarrow L_3 - L_2} \begin{cases} x+y+z=6, \\ \ \ y+z=5, \\ \ \ \ \ z=3. \end{cases}$$

- ⋆ échanger deux lignes,

$$L_i \leftrightarrow L_j$$

comme par exemple

$$\begin{cases} x+y+z=6, \\ \ \ \ \ z=3, \\ \ \ y+z=5, \end{cases} \xrightarrow{L_2 \leftrightarrow L_3} \begin{cases} x+y+z=6, \\ \ \ y+z=5, \\ \ \ \ \ z=3. \end{cases}$$

Ces transformations sont équivalentes à la multiplication à gauche (pré-multiplication) de la matrice $\mathbb{M} \in \mathscr{M}_{m,n}$ *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* $\mathbb{M} \in \mathscr{M}_{4,3}$ *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* $\mathbb{I}_4$ *:*

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

Méthode de Gauss

La méthode de GAUSS transforme un système linéaire quelconque en un système *échelonné* équivalent.
Soit $\mathbb{A} = (a_{ij})_{\substack{1 \le i \le n \\ 1 \le j \le p}}$ la matrice des coefficients du système (S) et $[\mathbb{A}|\mathbf{b}]$ 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 :

$$\begin{cases} x_1+2x_2+3x_3+4x_4=1 \\ 2x_1+3x_2+4x_3+x_4=2 \\ 3x_1+4x_2+x_3+2x_4=3 \\ 4x_1+x_2+2x_3+3x_4=4 \end{cases} \xrightarrow[\substack{L_3\leftarrow L_3-3L_1 \\ L_4\leftarrow L_4-4L_1}]{L_2\leftarrow L_2-2L_1 \\ \text{Étape 1}} \begin{cases} x_1+2x_2+3x_3+4x_4=1 \\ -x_2-2x_3-7x_4=0 \\ -2x_2-8x_3-10x_4=0 \\ -7x_2-10x_3-13x_4=0 \end{cases}$$

$$\xrightarrow[\substack{L_4\leftarrow L_4-7L_2}]{L_3\leftarrow L_3-2L_2 \\ \text{Étape 2}} \begin{cases} x_1+2x_2+3x_3+4x_4=1 \\ -x_2-2x_3-7x_4=0 \\ -4x_3+4x_4=0 \\ 4x_3+36x_4=0 \end{cases} \xrightarrow[\text{Étape 3}]{L_4\leftarrow L_4+L_3} \begin{cases} x_1+2x_2+3x_3+4x_4=1 \\ -x_2-2x_3-7x_4=0 \\ -4x_3+4x_4=0 \\ 40x_4=0 \end{cases}$$

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 :

$$[\mathbb{A}|\mathbf{b}] = \left(\begin{array}{cccc|c} 1&2&3&4&1 \\ 2&3&4&1&2 \\ 3&4&1&2&3 \\ 4&1&2&3&4 \end{array}\right) \xrightarrow[\substack{L_3\leftarrow L_3-3L_1 \\ L_4\leftarrow L_4-4L_1}]{L_2\leftarrow L_2-2L_1 \\ \text{Étape 1}} \left(\begin{array}{cccc|c} 1&2&3&4&1 \\ \boxed{0}&-1&-2&-7&0 \\ \boxed{0}&-2&-8&-10&0 \\ \boxed{0}&-7&-10&-13&0 \end{array}\right)$$

$$\xrightarrow[\substack{L_4\leftarrow L_4-7L_2}]{L_3\leftarrow L_3-2L_2 \\ \text{Étape 2}} \left(\begin{array}{cccc|c} 1&2&3&4&1 \\ 0&-1&-2&-7&0 \\ 0&\boxed{0}&-4&4&0 \\ 0&\boxed{0}&4&36&0 \end{array}\right) \xrightarrow[\text{Étape 3}]{L_4\leftarrow L_4+L_3} \left(\begin{array}{cccc|c} 1&2&3&4&1 \\ 0&-1&-2&-7&0 \\ 0&0&-4&4&0 \\ 0&0&\boxed{0}&40&0 \end{array}\right)$$

donc

$$x_4=0, \quad x_3=0, \quad x_2=0, \quad x_1=1.$$

EXEMPLE (SYSTÈME AVEC DES PARAMÈTRES)
Pour quelles valeurs de a et c le système linéaire suivant admet aucune, une seule ou une infinité de solutions ?

$$\begin{cases} x+5y+z=0, \\ x+6y-z=2, \\ 2x+ay+z=c. \end{cases}$$

Nous avons 3 équations donc il faut effectuer 2 étapes de la méthode de GAUSS :

$$\begin{cases} x+5y+z=0 \\ x+6y-z=2 \\ 2x+ay+z=c \end{cases} \xrightarrow[\substack{L_3\leftarrow L_3-2L_1}]{L_2\leftarrow L_2-L_1 \\ \text{Étape } j=1} \begin{cases} x+5y+z=0 \\ y-2z=2 \\ (a-10)y-z=c \end{cases} \xrightarrow[\text{Étape } j=2]{L_3\leftarrow L_3-(a-10)L_2} \begin{cases} x+5y+z=0 \\ y-2z=2 \\ (2a-21)z=c-2(a-10) \end{cases}$$

Étudions la dernière équation :

$$(2a-21)z=(c-2a+20)$$

- Si $a\neq\frac{21}{2}$ alors $z=\frac{c-2a+20}{2a-21}$ et on trouve y puis x en remontant : il existe une et une seule solution ;
- si $a=\frac{21}{2}$ alors
  - si $c-2a+20=0$ (*i.e.* $c=1$), alors $z=\kappa\in\mathbb{R}$ et on trouve y puis x en remontant : il existe une infinité de solutions ;
  - si $c-2a+20\neq 0$ (*i.e.* $c\neq 1$), alors il n'y a aucune solution.

Variante 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.

La méthode de GAUSS-JORDAN comporte n étapes : à chaque étape j on fait apparaître des 0 sur la colonne j pour les lignes $i\neq j$ par des opérations élémentaires sur les lignes.

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 fait apparaître des 0 sur la colonne j pour les lignes $i \neq j$ (*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, 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.

$$[\mathbb{A}|\mathbf{b}] = \left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 2 & 3 & 4 & 1 & 2 \\ 3 & 4 & 1 & 2 & 3 \\ 4 & 1 & 2 & 3 & 4 \end{array}\right) \xrightarrow[\substack{L_3 \leftarrow L_3 - 3L_1 \\ L_4 \leftarrow L_4 - 4L_1}]{L_2 \leftarrow L_2 - 2L_1} \text{Étape 1} \left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ \boxed{0} & -1 & -2 & -7 & 0 \\ \boxed{0} & -2 & -8 & -10 & 0 \\ \boxed{0} & -7 & -10 & -13 & 0 \end{array}\right) \xrightarrow[\substack{L_3 \leftarrow L_3 - 2L_2 \\ L_4 \leftarrow L_4 - 7L_2}]{L_1 \leftarrow L_1 + 2L_2} \text{Étape 2} \left(\begin{array}{cccc|c} 1 & \boxed{0} & -1 & -10 & 1 \\ 0 & -1 & -2 & -7 & 0 \\ 0 & \boxed{0} & -4 & 4 & 0 \\ 0 & \boxed{0} & 4 & 36 & 0 \end{array}\right)$$

$$\xrightarrow[\substack{L_2 \leftarrow L_2 - L_3/2 \\ L_4 \leftarrow L_4 + L_3}]{L_1 \leftarrow L_1 - L_3/4} \text{Étape 3} \left(\begin{array}{cccc|c} 1 & 0 & \boxed{0} & 4 & 1 \\ 0 & -1 & \boxed{0} & -7 & 0 \\ 0 & 0 & -4 & 4 & 0 \\ 0 & 0 & \boxed{0} & 40 & 0 \end{array}\right) \xrightarrow[\substack{L_2 \leftarrow L_2 + 9L_4/40 \\ L_3 \leftarrow L_3 + 4L_4/40}]{L_1 \leftarrow L_1 + 11L_4/40} \text{Étape 4} \left(\begin{array}{cccc|c} 1 & 0 & 0 & \boxed{0} & 1 \\ 0 & -1 & 0 & \boxed{0} & 0 \\ 0 & 0 & -4 & \boxed{0} & 0 \\ 0 & 0 & 0 & 40 & 0 \end{array}\right)$$

donc

$$x_1 = 1, \quad x_2 = 0, \quad x_3 = 0, \quad x_4 = 0.$$

Rang d'un système linéaire

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 1.20
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 :

$$① \begin{cases} x + y = 1 \\ x - y = 1 \end{cases} \qquad ② \begin{cases} x + y = 1 \\ 2x + 2y = 2 \end{cases} \qquad ③ \begin{cases} x + y = 1 \\ 2x + 2y = 1 \end{cases}$$

Les matrices augmentées associées à chaque système sont

$$① [\mathbb{A}|\mathbf{b}] = \left(\begin{array}{cc|c} 1 & 1 & 1 \\ 1 & -1 & 1 \end{array}\right) \qquad ② [\mathbb{A}|\mathbf{b}] = \left(\begin{array}{cc|c} 1 & 1 & 1 \\ 2 & 2 & 2 \end{array}\right) \qquad ③ [\mathbb{A}|\mathbf{b}] = \left(\begin{array}{cc|c} 1 & 1 & 1 \\ 2 & 2 & 1 \end{array}\right)$$

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 $\mathrm{rg}(\mathbb{A})=1$ et $\mathrm{rg}([\mathbb{A}|\mathbf{b}])=2$ 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.

Système de Cramer et méthode de Cramer

Un SYSTÈME est dit DE CRAMER s'il a une solution, et une seule.

Propriété 1.21
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. $\mathrm{rg}(\mathbb{A})=n$;
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

$$x_j=\frac{\det(\mathbb{A}_j)}{\det(\mathbb{A})},\qquad 1\le j\le n$$

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

$$\mathbb{A}=\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix},\qquad \det(\mathbb{A})=a_{11}a_{22}-a_{12}a_{21},$$
$$\mathbb{A}_1=\begin{pmatrix} b_1 & a_{12} \\ b_2 & a_{22} \end{pmatrix},\qquad \det(\mathbb{A}_1)=b_1a_{22}-a_{12}b_2,$$
$$\mathbb{A}_2=\begin{pmatrix} a_{11} & b_1 \\ a_{21} & b_2 \end{pmatrix},\qquad \det(\mathbb{A}_2)=a_{11}b_2-b_1a_{21},$$

donc

$$x_1=\frac{b_1a_{22}-a_{12}b_2}{a_{11}a_{22}-a_{12}a_{21}},\qquad x_2=\frac{a_{11}b_2-b_1a_{21}}{a_{11}a_{22}-a_{12}a_{21}}.$$

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

$$\mathbb{A}=\begin{pmatrix} 1 & -1 & 2 \\ 2 & 1 & 0 \\ 3 & 2 & 0 \end{pmatrix},\qquad \det(\mathbb{A})=2,$$
$$\mathbb{A}_1=\begin{pmatrix} 2 & -1 & 2 \\ -1 & 1 & 0 \\ 1 & 2 & 0 \end{pmatrix},\qquad \det(\mathbb{A}_1)=-6,$$

$$\mathbb{A}_2 = \begin{pmatrix} 1 & 2 & 2 \\ 2 & -1 & 0 \\ 3 & 1 & 0 \end{pmatrix}, \qquad \det(\mathbb{A}_2) = 10,$$

$$\mathbb{A}_3 = \begin{pmatrix} 1 & -1 & 2 \\ 2 & 1 & -1 \\ 3 & 2 & 1 \end{pmatrix}, \qquad \det(\mathbb{A}_3) = 10,$$

donc

$$x = \frac{-6}{2} = -3, \qquad y = \frac{10}{2} = 5, \qquad z = \frac{10}{2} = 5.$$

Définition 1.22 (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 $\begin{pmatrix} d & -c \\ -b & a \end{pmatrix}$.

1.3.1. Calcul de la matrice inverse

$\mathbb{A}$ 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 $(1,0,0,\ldots,0)$, $(0,1,0,\ldots,0)$, …, $(0,0,0,\ldots,1)$. 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 $\mathbb{A} = \begin{pmatrix} 2 & 0 \\ 2 & 2 \end{pmatrix}$. Comme $\det(\mathbb{A}) = 4 \neq 0$ la matrice est inversible.

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

$$\mathbb{A}^{-1} = \frac{1}{4}\begin{pmatrix} 2 & 0 \\ -2 & 2 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & 0 \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

Deuxième méthode : on parvient au même résultat par transformations élémentaires :

$$[\mathbb{A}|\mathbb{I}_2] = \left(\begin{array}{cc|cc} 2 & 0 & 1 & 0 \\ 2 & 2 & 0 & 1 \end{array}\right) \xrightarrow{L_2 \leftarrow L_2 - L_1} \left(\begin{array}{cc|cc} 2 & 0 & 1 & 0 \\ 0 & 2 & -1 & 1 \end{array}\right)$$
$$\xrightarrow[L_2 \leftarrow \frac{1}{2}L_2]{L_1 \leftarrow \frac{1}{2}L_1} \left(\begin{array}{cc|cc} 1 & 0 & \frac{1}{2} & 0 \\ 0 & 1 & -\frac{1}{2} & \frac{1}{2} \end{array}\right)$$

EXEMPLE
Soit $\mathbb{A} = \begin{pmatrix} 2 & 1 \\ 2 & 2 \end{pmatrix}$. Comme $\det(\mathbb{A}) = 2 \neq 0$ la matrice est inversible.

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

$$\mathbb{A}^{-1} = \frac{1}{2}\begin{pmatrix} 2 & -1 \\ -2 & 2 \end{pmatrix} = \begin{pmatrix} 1 & -\frac{1}{2} \\ -1 & 1 \end{pmatrix}.$$

Deuxième méthode : on parvient au même résultat par transformations élémentaires :

$$[\mathbb{A}|\mathbb{I}_2] = \left(\begin{array}{cc|cc} 2 & 1 & 1 & 0 \\ 2 & 2 & 0 & 1 \end{array}\right) \xrightarrow{L_2 \leftarrow L_2 - L_1} \left(\begin{array}{cc|cc} 2 & 1 & 1 & 0 \\ 0 & 1 & -1 & 1 \end{array}\right)$$
$$\xrightarrow{L_1 \leftarrow L_1 - L_2} \left(\begin{array}{cc|cc} 2 & 0 & 2 & -1 \\ 0 & 1 & -1 & 1 \end{array}\right)$$
$$\xrightarrow{L_1 \leftarrow \frac{1}{2}L_1} \left(\begin{array}{cc|cc} 1 & 0 & 1 & -\frac{1}{2} \\ 0 & 1 & -1 & 1 \end{array}\right)$$

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 :

$$[\mathbb{A}|\mathbb{I}_2] = \left(\begin{array}{cc|cc} a & b & 1 & 0 \\ c & d & 0 & 1 \end{array}\right) \xrightarrow{L_2 \leftarrow L_2 - \frac{c}{a}L_1} \left(\begin{array}{cc|cc} a & b & 1 & 0 \\ 0 & d - \frac{c}{a}b & -\frac{c}{a} & 1 \end{array}\right)$$
$$\xrightarrow[L_2 = L_1 - \frac{ab}{ad-bc}L_2]{L_1 \leftarrow L_1 - \frac{b}{d-\frac{c}{a}b}} \left(\begin{array}{cc|cc} a & 0 & 1 + \frac{bc}{ad-bc} & -\frac{ab}{ad-bc} \\ 0 & d - \frac{c}{a}b & -\frac{c}{a} & 1 \end{array}\right)$$
$$\xrightarrow[L_2 \leftarrow \frac{1}{d-\frac{c}{a}b}L_2 = \frac{a}{ad-cb}L_2]{L_1 \leftarrow \frac{1}{a}L_1} \left(\begin{array}{cc|cc} 1 & 0 & \frac{1}{a} + \frac{bc}{a(ad-bc)} & -\frac{ab}{ad-bc} \\ 0 & 1 & -\frac{c}{ad-cb} & \frac{a}{ad-cb} \end{array}\right) = \left(\begin{array}{cc|cc} 1 & 0 & \frac{d}{ad-bc} & -\frac{b}{ad-bc} \\ 0 & 1 & -\frac{c}{ad-cb} & \frac{a}{ad-cb} \end{array}\right)$$

EXEMPLE
Calculer l'inverse de la matrice

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

Première méthode.

1. On calcule la matrice des cofacteurs des éléments de $\mathbb{A}$, appelée comatrice de $\mathbb{A}$:

$$\text{comatrice} = \begin{pmatrix} (-1)^{1+1}\begin{vmatrix} 1 & 1 \\ -1 & 1 \end{vmatrix} & (-1)^{1+2}\begin{vmatrix} -1 & 1 \\ 1 & 1 \end{vmatrix} & (-1)^{1+3}\begin{vmatrix} -1 & 1 \\ 1 & -1 \end{vmatrix} \\ (-1)^{2+1}\begin{vmatrix} 1 & -1 \\ -1 & 1 \end{vmatrix} & (-1)^{2+2}\begin{vmatrix} 1 & -1 \\ 1 & 1 \end{vmatrix} & (-1)^{2+3}\begin{vmatrix} 1 & 1 \\ 1 & -1 \end{vmatrix} \\ (-1)^{3+1}\begin{vmatrix} 1 & -1 \\ 1 & 1 \end{vmatrix} & (-1)^{3+2}\begin{vmatrix} 1 & -1 \\ -1 & 1 \end{vmatrix} & (-1)^{3+3}\begin{vmatrix} 1 & 1 \\ -1 & 1 \end{vmatrix} \end{pmatrix} = \begin{pmatrix} 2 & 2 & 0 \\ 0 & 2 & 2 \\ 2 & 0 & 2 \end{pmatrix};$$

2. on transpose la comatrice de $\mathbb{A}$:

$$\text{comatrice}^T = \begin{pmatrix} 2 & 0 & 2 \\ 2 & 2 & 0 \\ 0 & 2 & 2 \end{pmatrix};$$

3. 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.

$$[\mathbb{A}|\mathbb{I}_3] = \left(\begin{array}{ccc|ccc} 1 & 1 & -1 & 1 & 0 & 0 \\ -1 & 1 & 1 & 0 & 1 & 0 \\ 1 & -1 & 1 & 0 & 0 & 1 \end{array}\right) \xrightarrow[L_3 \leftarrow L_3 - L_1]{L_2 \leftarrow L_2 + L_1} \left(\begin{array}{ccc|ccc} 1 & 1 & -1 & 1 & 0 & 0 \\ \boxed{0} & 2 & 0 & 1 & 1 & 0 \\ \boxed{0} & -2 & 2 & -1 & 0 & 1 \end{array}\right)$$

$$\xrightarrow{L_2 \leftarrow L_2/2} \left(\begin{array}{ccc|ccc} 1 & 1 & -1 & 1 & 0 & 0 \\ 0 & \boxed{1} & 0 & 1/2 & 1/2 & 0 \\ 0 & -2 & 2 & -1 & 0 & 1 \end{array}\right) \xrightarrow[L_3 \leftarrow L_3 + 2L_2]{L_1 \leftarrow L_1 - L_2} \left(\begin{array}{ccc|ccc} 1 & \boxed{0} & -1 & 1/2 & -1/2 & 0 \\ 0 & 1 & 0 & 1/2 & 1/2 & 0 \\ 0 & \boxed{0} & 2 & 0 & 1 & 1 \end{array}\right)$$

$$\xrightarrow{L_3 \leftarrow L_3/2} \left(\begin{array}{ccc|ccc} 1 & 0 & -1 & 1/2 & -1/2 & 0 \\ 0 & 1 & 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & \boxed{1} & 0 & 1/2 & 1/2 \end{array}\right) \xrightarrow{L_1 \leftarrow L_1 + L_3} \left(\begin{array}{ccc|ccc} 1 & 0 & \boxed{0} & 1/2 & 0 & 1/2 \\ 0 & 1 & \boxed{0} & 1/2 & 1/2 & 0 \\ 0 & 0 & 1 & 0 & 1/2 & 1/2 \end{array}\right) = [\mathbb{I}_3|\mathbb{A}^{-1}].$$

1.3.2. Système sur-déterminé

Si le système (S) a n équations et m inconnues avec $n > m$, on dit que le système est sur-déterminé. On considère alors (S') un sous-système carré d'ordre m qu'on peut résoudre par exemple par la méthode du pivot de Gauss. Parmi les solutions de ce système carré, on cherchera celles qui vérifient les équations de (S) qui n'apparaissent pas dans (S').

EXEMPLE
Soit les systèmes linéaires de $n = 3$ équations et $m = 2$ inconnues

$$(S_1) \quad \begin{cases} x + y = 2 \\ x + 2y = 3 \\ x + 3y = 4 \end{cases} \qquad (S_2) \quad \begin{cases} x + y = 2 \\ x + 2y = 3 \\ x + 3y = 0 \end{cases}$$

Prenons comme sous-système carré d'ordre $m = 2$ celui constitué des deux premières équations et résolvons-le :

$$(S') \quad \begin{cases} x + y = 2 \\ x + 2y = 3 \end{cases} \xrightarrow{L_2 \leftarrow L_2 - L_1} \begin{cases} x + y = 2 \\ y = 1 \end{cases}$$

Ce système admet une seule solution : $x = y = 1$.
On vérifie si cette solution satisfait l'équation de (S_1) qui n'apparaît pas dans (S') :

$$x + 3y = 1 + 3 = 4$$

donc $x = y = 1$ est l'unique solution de (S_1).
On vérifie si cette solution satisfait l'équation de (S_2) qui n'apparaît pas dans (S') :

$$x + 3y = 1 + 3 = 4 \neq 0$$

donc (S_2) n'admet pas de solution.

1.3.3. Système sous-déterminé

Un système est sous-déterminé si, après échelonnage, le nombre d'équations significatives est inférieur au nombre d'inconnues.

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.
Toutes ces contraintes dépendent linéairement des coefficients stœchiométriques, ce qui amène tout naturellement à l'écriture d'un système linéaire.
Typiquement on aura n inconnues mais seulement $n - 1$ 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 les coefficients trouvés équilibrent la réaction, alors tous les multiples entiers de ces coefficients équilibrent aussi la réaction.

EXEMPLE
Si on mélange de la soude caustique et de l'acide sulfurique, on obtient du sulfate de sodium et de l'eau :

$$\underbrace{?NaOH}_{\text{soude caustique}} + \underbrace{?H_2SO_4}_{\text{acide sulfurique}} \rightarrow \underbrace{?Na_2SO_4}_{\text{sulfate de sodium}} + ?H_2O$$

Pour que cette réaction ait lieu, il faut que tous les atomes (par exemple de sodium) qui sont à gauche se retrouvent à droite et vice-versa.

$$\underbrace{x\,NaOH}_{\text{soude caustique}} + \underbrace{y\,H_2SO_4}_{\text{acide sulfurique}} \rightarrow \underbrace{z\,Na_2SO_4}_{\text{sulfate de sodium}} + w\,H_2O$$

On voit bien qu'il nous faut au moins 2 molécules de NaOH à gauche pour tomber sur le Na_2 de droite. On pose alors $x = 2$ (mieux, un multiple de 2) et $z = 1$:

$$2NaOH + yH_2SO_4 \rightarrow Na_2SO_4 + wH_2O$$

Le 2OH à gauche venant de la soude et la yH_2 venant de l'acide sulfurique se combinent pour donner wH_2O. On peut alors poser $y = 1$ et $w = 2$:

$$2NaOH + H_2SO_4 \rightarrow Na_2SO_4 + 2H_2O$$

Le SO_4 se trouve bien à gauche et à droite et l'équation est alors équilibrée.
En système cela devient

$$\begin{cases} x = 2z & [Na] \\ x + 2y = 2w & [H] \\ x + 4y = 4z + w & [O] \\ y = z & [S] \end{cases}$$

On trouve $z = y = \kappa$, $x = 2\kappa$ et $w = \kappa$ et l'équation est alors équilibrée. On peut alors poser $\kappa = 1$.

EXEMPLE
Considérons la réaction

$$H_2 + O_2 \rightarrow H_2O.$$

Notons x_1, x_2 et x_3 les coefficients stœchiométriques

$$x_1 H_2 + x_2 O_2 \rightarrow x_3 H_2 O.$$

Les contraintes sont :

- la conservation du nombre d'atomes d'hydrogène : $2x_1 = 2x_3$,
- la conservation du nombre d'atomes d'oxygène : $2x_2 = x_3$.

On note qu'on a 3 inconnues mais seulement 2 équations linéairement indépendantes.
Pour résoudre le problème sans paramètres, fixons arbitrairement un des coefficients, par exemple $x_3 = 1$. On doit alors résoudre le système linéaire

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

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

$$2H_2 + 1O_2 \rightarrow 2H_2O.$$

1.3.4. Conclusion sur les systèmes rectangulaires

Théorème 1.23
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 $\mathbb{A} \in \mathscr{M}_{n,p}$ la matrice des coefficients du système (S). Alors

$$0 \leq \text{rg}(\mathbb{A}) \leq \min\{n, p\}$$
$$\text{rg}(\mathbb{A}) \leq \text{rg}([\mathbb{A}|\mathbf{b}]) \leq \min\{n, p+1\}.$$

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 cherches 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 :

$$\mathbb{A}=\begin{pmatrix}1&1&3&1\\1&3&2&4\\2&0&1&-1\end{pmatrix}\xrightarrow[L_3\leftarrow L_3-2L_1]{L_2\leftarrow L_2-L_1}\begin{pmatrix}1&1&3&1\\0&2&-1&3\\0&-2&-5&-3\end{pmatrix}\xrightarrow{L_3\leftarrow L_3+L_2}\begin{pmatrix}1&1&3&1\\0&2&-1&3\\0&0&-6&0\end{pmatrix}$$

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.

1.4. Valeurs propres et vecteurs propres

Une matrice peut être représentée par ses valeurs propres et ses vecteurs propres. Cette représentation est appelée *décomposition en valeurs propres*. Pour une matrice $\mathbb{A}$ donnée, la notion clé est la résolution de l'équation

$$\mathbb{A}\mathbf{v}=\lambda\mathbf{v}$$

où $\mathbf{v}$ est un vecteur propre de la matrice $\mathbb{A}$ et λ est la valeur propre correspondante. λ et $\mathbf{v}$ sont appelés un couple valeur propre-vecteur propre.
Dans cette partie, nous allons explorer la relation entre une matrice et sa décomposition en vecteurs propres.

1.4.1. Produit matrice-vecteur et lien avec la décomposition en valeurs propres

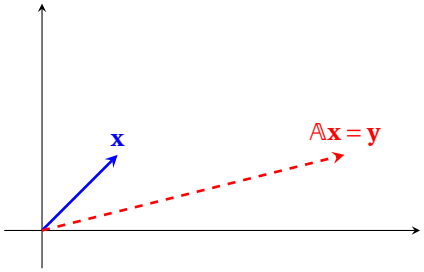
Lorsque on multiplie une matrice et un vecteur, le résultat est un autre vecteur. De cette façon, la matrice $\mathbb{A}$ peut être considérée comme une transformation qui transforme un vecteur $\mathbf{x}$ en un vecteur $\mathbf{y}$.
Par exemple, si $\mathbf{x}$ est un vecteur de $\mathbb{R}^2$, on peut visualiser cette multiplication comme suit.

Dans le dessin ci-contre,

$$\mathbf{x} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \mathbb{A} = \begin{pmatrix} 0 & 4 \\ 1 & 0 \end{pmatrix},$$

ainsi

$$\mathbf{y} = \mathbb{A}\mathbf{x} = \begin{pmatrix} 4 \\ 1 \end{pmatrix}.$$



Lorsque on multiplie une matrice et son vecteur propre, c'est comme si le vecteur propre venait d'être multiplié par un nombre mais la direction n'est pas modifiée : il est juste mis à l'échelle et le facteur de mise à l'échelle est la valeur propre.

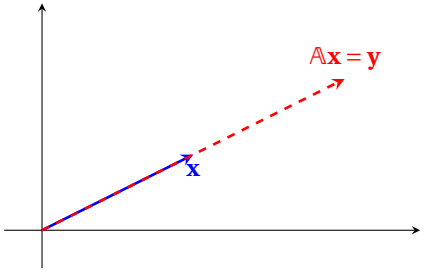
Dans le dessin ci-contre on a la même matrice mais

$$\mathbf{x} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

ainsi

$$\mathbf{y} = \mathbb{A}\mathbf{x} = \begin{pmatrix} 4 \\ 2 \end{pmatrix} = 2\mathbf{x},$$

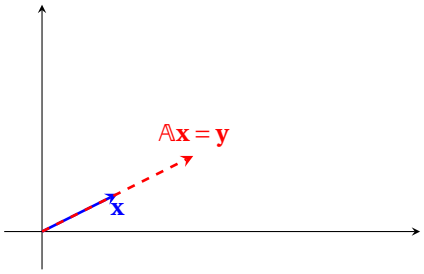
on dit que $\lambda = 2$ est une valeur propre pour $\mathbb{A}$ et $\mathbf{x} = (2,1)$ une vecteur propre associée à cette valeur propre.



Bien sûr,

$$\mathbf{x} = \begin{pmatrix} 1 \\ \frac{1}{2} \end{pmatrix}$$

est encore une vecteur propre associée à la valeur propre $\lambda = 2$ comme on voit sur la figure ci-contre.



1.4.2. Définitions et propriétés

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 λ, 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 $\varrho(\mathbb{A})$.
- ⋆ On dit que deux matrices $\mathbb{A}$ et $\mathbb{B}$ carrées d'ordre n sont semblables s'il existe une matrice $\mathbb{P}$ carrées d'ordre n inversible telle que $\mathbb{A} = \mathbb{P}^{-1}\mathbb{B}\mathbb{P}$. On peut démonter que
  - ⋆ $p_{\mathbb{A}}(\lambda) = p_{\mathbb{B}}(\lambda)$;
  - ⋆ $\sigma(\mathbb{A}) = \sigma(\mathbb{B})$.
- ⋆ On appelle VECTEUR PROPRE DE $\mathbb{A}$ ASSOCIÉ À LA VALEUR PROPRE λ 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 λ 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 démonter que

1. $\det(\mathbb{A}) = \prod_{i=1}^{n} \lambda_i$, (donc $\det(\mathbb{A}) = 0$ ssi il existe une valeur propre nulle) ;
2. $\mathrm{tr}(\mathbb{A}) = \sum_{i=1}^{n} \lambda_i$;
3. $\sigma(\mathbb{A}^T) = \sigma(\mathbb{A})$ et $\sigma(\mathbb{A}^H) = \sigma(\mathbb{A})$;
4. λ est une valeur propre de $\mathbb{A} \in \mathbb{C}^{n\times n} \iff \bar{\lambda}$ est une valeur propre de $\mathbb{A}^H$.

Une matrice carrée $\mathbb{A}$ d'ordre n est diagonalisable si elle est semblable à une matrice diagonale. On peut démonter 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 $\mathbb{D} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ 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 $\mathbb{D}^p = \mathrm{diag}(\lambda_1^p, \lambda_2^p, \ldots, \lambda_n^p)$;
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 $\mathbb{R}$.

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

$$\begin{aligned} p(\lambda) = \det(\mathbb{A} - \lambda\mathbb{I}) &= \det\begin{pmatrix} 1-\lambda & -1 & 0 \\ -1 & 2-\lambda & -1 \\ 0 & -1 & 1-\lambda \end{pmatrix} \\ &= (1-\lambda)\det\begin{pmatrix} 2-\lambda & -1 \\ -1 & 1-\lambda \end{pmatrix} - (-1)\det\begin{pmatrix} -1 & -1 \\ 0 & 1-\lambda \end{pmatrix} \\ &= (1-\lambda)\Big((2-\lambda)(1-\lambda)-1\Big) - (1-\lambda) = (1-\lambda)\Big((2-\lambda)(1-\lambda)-2\Big) = (1-\lambda)\Big(-3\lambda+\lambda^2\Big) = \lambda(1-\lambda)(\lambda-3) \end{aligned}$$

  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 valeurs propre λ_1.
On cherche $\mathbf{x}$ tel que

$$(\mathbb{A}-\lambda_1\mathbb{I})\mathbf{x}=\mathbf{0} \qquad \text{c'est-à-dire} \qquad \begin{pmatrix}1-\lambda_1 & -1 & 0\\ -1 & 2-\lambda_1 & -1\\ 0 & -1 & 1-\lambda_1\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=\begin{pmatrix}0\\0\\0\end{pmatrix}.$$

En utilisant la méthode de Gauss (le système étant homogène, on n'écrit pas le second membre) on a

$$\begin{pmatrix}1 & -1 & 0\\ -1 & 2 & -1\\ 0 & -1 & 1\end{pmatrix}\xrightarrow[\substack{L_3\leftarrow L_3\\ \text{Étape 1}}]{L_2\leftarrow L_2+L_1}\begin{pmatrix}1 & -1 & 0\\ 0 & 1 & -1\\ 0 & -1 & 1\end{pmatrix}\xrightarrow[\text{Étape 2}]{L_3\leftarrow L_3+L_2}\begin{pmatrix}1 & -1 & 0\\ 0 & 1 & -1\\ 0 & 0 & 0\end{pmatrix}$$

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

$$\mathbf{x}=\kappa\begin{pmatrix}1\\1\\1\end{pmatrix}.$$

Pour faire simple, on choisira $\kappa=1$.

⋆ Calcul des vecteurs propres associés à la valeurs propre λ_2.
On cherche $\mathbf{x}$ tel que

$$(\mathbb{A}-\lambda_2\mathbb{I})\mathbf{x}=\mathbf{0} \qquad \text{c'est-à-dire} \qquad \begin{pmatrix}1-\lambda_2 & -1 & 0\\ -1 & 2-\lambda_2 & -1\\ 0 & -1 & 1-\lambda_2\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=\begin{pmatrix}0\\0\\0\end{pmatrix}.$$

En utilisant la méthode de Gauss on a

$$\begin{pmatrix}0 & -1 & 0\\ -1 & 1 & -1\\ 0 & -1 & 0\end{pmatrix}\xrightarrow[\text{Étape 1}]{L_2\leftarrow L_1}\begin{pmatrix}-1 & 1 & -1\\ 0 & -1 & 0\\ 0 & -1 & 0\end{pmatrix}\xrightarrow[\text{Étape 2}]{L_3\leftarrow L_3-L_2}\begin{pmatrix}-1 & 1 & -1\\ 0 & -1 & 0\\ 0 & 0 & 0\end{pmatrix}$$

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

$$\mathbf{x}=\kappa\begin{pmatrix}-1\\0\\1\end{pmatrix}.$$

Pour faire simple, on choisira $\kappa=1$.

⋆ Calcul des vecteurs propres associés à la valeurs propre λ_3.
On cherche $\mathbf{x}$ tel que

$$(\mathbb{A}-\lambda_3\mathbb{I})\mathbf{x}=\mathbf{0} \qquad \text{c'est-à-dire} \qquad \begin{pmatrix}1-\lambda_3 & -1 & 0\\ -1 & 2-\lambda_3 & -1\\ 0 & -1 & 1-\lambda_3\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=\begin{pmatrix}0\\0\\0\end{pmatrix}.$$

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

$$\mathbf{x} = \kappa \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}.$$

Pour faire simple, on choisira $\kappa = 1$.

⋆ *Diagonalisation*
On peut alors écrire les valeurs propres et les vecteurs propres dans deux matrices

$$\mathbb{D} = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{pmatrix} \quad \text{et} \quad \mathbb{P} = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & -2 \\ 1 & -1 & 1 \end{pmatrix}$$

et vérifier que $\mathbb{A} = \mathbb{P}\mathbb{D}\mathbb{P}^{-1}$, c'est-à-dire que $\mathbb{A}\mathbb{P} = \mathbb{P}\mathbb{D}$:

$$\mathbb{A}\mathbb{P} = \begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & -2 \\ 1 & -1 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 3 \\ 0 & 0 & -6 \\ 0 & -1 & 3 \end{pmatrix}$$

$$\mathbb{P}\mathbb{D} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & -2 \\ 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 3 \\ 0 & 0 & -6 \\ 0 & -1 & 3 \end{pmatrix}$$

Une matrice carrée $\mathbb{A}$ d'ordre n n'est pas toujours diagonalisable. En revanche, elle est toujours trigonalisable sur $\mathbb{C}$, *i.e.* elle est semblable à une matrice triangulaire et l'on a le résultat suivant :

Proposition 1.24 (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 λ_i les valeurs propres de $\mathbb{A}$. Les matrices $\mathbb{U}$ et $\mathbb{T}$ ne sont pas forcement uniques.

On peut démonter que

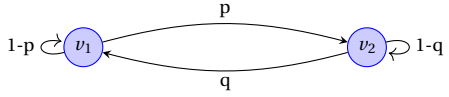
1. si $\mathbb{A}$ est hermitienne alors $\mathbb{T}$ est toujours une matrice diagonale et les colonnes de $\mathbb{U}$ sont les vecteurs propres de $\mathbb{A}$;
2. si de plus $\mathbb{A}$ est normale alors on a la décomposition spectrale $\mathbb{A} = \mathbb{U}\mathbb{T}\mathbb{U}^H = \sum_{i=1}^n \lambda_i \mathbf{u}_i \mathbf{u}_i^H$

1.4.3. Applications

Marche aléatoire entre deux états (chaîne de Markov)

Lorsqu'un système n'ayant que deux états possibles 1 et 2 évolue par étapes successives aléatoires et indépendantes, on dit qu'il suit une marche aléatoire entre ses deux états. Soit p la probabilité qu'il passe de 1 à 2 et q la probabilité qu'il passe de 2 à 1. On peut alors lui associer :

⋆ un graphe probabiliste qui schématise les échanges entre 1 et 2 par des arêtes orientées, pondérées par les probabilités de passer d'un état à l'autre ou de rester au même état,



⋆ une matrice de transition $\mathbb{T}$ carrée d'ordre 2 telle que le coefficient t_{ij} est égal à la probabilité

- ⋆ de passer de l'état j à l'état i lorsque $i \neq j$;
- ⋆ de rester à l'état i lorsque $i = j$.

$$\mathbb{T} = \begin{pmatrix} 1-p & q \\ p & 1-q \end{pmatrix}$$

Une matrice de transition est dite stochastique : ses coefficients appartiennent à l'intervalle $[0;1]$ et la somme des coefficients de chacune de ses colonnes est égale à 1.

Pour $n \in \mathbb{N}$, on note

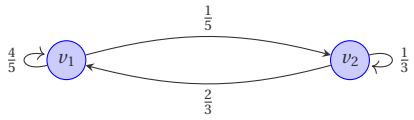
- ⋆ l'événement A_n : « le système est dans l'état A à l'étape n » ;
- ⋆ l'événement B_n : « le système est dans l'état B à l'étape n » ;
- ⋆ les probabilités $a_n = P(A_n)$ et $b_n = P(B_n)$ telles que $a_n + b_n = 1$.

Le vecteur $\mathbf{u}^{(n)} = \begin{pmatrix} a_n \\ b_n \end{pmatrix}$ est appelée la **répartition de probabilité à l'étape** n et l'on a

$$\mathbf{u}^{(n+1)} = \mathbb{T}\mathbf{u}^{(n)} = \mathbb{T}^n\mathbf{u}^{(0)}$$

EXEMPLE

Akwa, un chien ayant une puce, rencontre Bali, un autre chien. Chaque seconde, la puce reste sur un chien ou va sur l'autre. On a un système à deux états : l'état 1 (la puce est sur Akwa) et l'état 2 (la puce est sur Bali) dont l'évolution est une marche aléatoire entre ces deux états.
Supposons que chaque seconde soit la puce va d'Akwa (état 1) à Bali (état 2) une fois sur cinq, soit elle va de Bali à Akwa deux fois sur trois, soit elle reste sur le même chien. Alors, la marche aléatoire a pour graphe



et matrice de transition :

$$\mathbb{T} = \begin{pmatrix} 4/5 & 2/3 \\ 1/5 & 1/3 \end{pmatrix}$$

Initialement, la puce est sur Akwa donc $\mathbf{u}^{(0)} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$.
Après une seconde, la répartition de probabilité est

$$\mathbf{u}^{(1)} = \mathbb{T}\mathbf{u}^{(0)} = \begin{pmatrix} 4/5 & 2/3 \\ 1/5 & 1/3 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 4/5 \\ 1/5 \end{pmatrix}$$

Après deux secondes, la répartition de probabilité est

$$\mathbf{u}^{(2)} = \mathbb{T}\mathbf{u}^{(1)} = \begin{pmatrix} 4/5 & 2/3 \\ 1/5 & 1/3 \end{pmatrix}\begin{pmatrix} 4/5 \\ 1/5 \end{pmatrix} = \begin{pmatrix} 58/75 \\ 17/75 \end{pmatrix}$$

Pour calculer $\lim_{n\to+\infty} \mathbf{u}^{(n+1)}$, il faut calculer $\lim_{n\to+\infty} \mathbb{T}\mathbf{u}^{(n)} = \mathbb{T}(\lim_{n\to+\infty} \mathbf{u}^{(n)})$.
Supposons qu'une telle limite existe et notons-la $\mathbf{u}$, alors $\mathbf{u}$ vérifie $\mathbf{u} = \mathbb{T}\mathbf{u}$, autrement dit $\lambda = 1$ est une valeur propre de $\mathbb{T}$ et $\mathbf{u}$ est le vecteur propre unitaire correspondant.
Le vecteur propre associé à $\lambda = 1$ est donné par

$$\begin{pmatrix} -\frac{1}{5} & \frac{2}{3} \\ \frac{1}{5} & -\frac{2}{3} \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

d'où $3y_1 = 10y_2$ soit encore $\mathbf{x} = (10\kappa, 3\kappa)^T$. La distribution normalisée est donnée par les composantes du vecteur propre unitaire correspondant, c'est-à-dire

$$\mathbf{x} = \frac{1}{10\kappa + 3\kappa}\begin{pmatrix} 10\kappa \\ 3\kappa \end{pmatrix} = \begin{pmatrix} \frac{10}{13} \\ \frac{3}{13} \end{pmatrix}.$$

Cela signifie que

$$\lim_{n\to+\infty} a_n = \frac{10}{13}, \qquad \lim_{n\to+\infty} b_n = \frac{3}{13}.$$

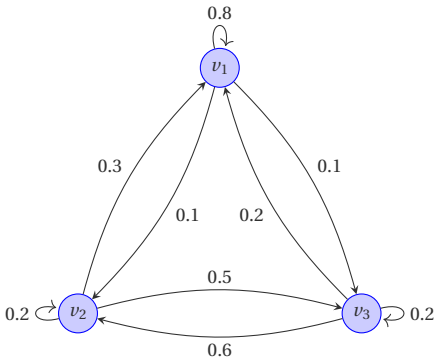
On généralise ces définitions et ces propriétés à des marches aléatoires entre trois états ou plus. Ainsi, le coefficient t_{ij} de la matrice de transition $\mathbb{T}$ est égal à la probabilité de passer de l'état j à l'état i.

EXEMPLE

v_1, v_2 et v_3 sont trois villes. Des trafiquants de drogue prennent leur marchandise le matin dans n'importe laquelle de ces villes pour l'apporter le soir dans n'importe quelle autre. On notera p_{ij} la probabilité qu'une marchandise prise le matin dans la ville v_j soit rendue le soir dans la ville v_i. On construit ainsi la matrice $\mathbb{P} \in \mathscr{M}_3([0;1])$, appelée *matrice de transition de la chaîne de* MARKOV. On remarque que la somme des composantes de chaque vecteur colonne est égale à 1.
Supposons que $\mathbb{A}$ soit connue et vaille

$$\mathbb{A} = \begin{pmatrix} 0.8 & 0.3 & 0.2 \\ 0.1 & 0.2 & 0.6 \\ 0.1 & 0.5 & 0.2 \end{pmatrix}.$$

Les trafiquants se promenant de ville en ville, il peut être utile de visualiser leurs déplacements par le diagramme de transition suivant :



On notera $x_i^{(k)}$ la proportion de trafiquants qui se trouvent au matin du jour k dans la ville v_i. On montre que le vecteur $\mathbf{x}^{(k)} = (x_1^{(k)}, x_2^{(k)}, x_3^{(k)})^T$ vérifie la relation

$$\mathbf{x}^{(k+1)} = \mathbb{A}\mathbf{x}^{(k)}$$

et donc par une récurrence immédiate

$$\mathbf{x}^{(k)} = \mathbb{A}^k \mathbf{x}^{(0)}.$$

Supposons que le chef de la mafia locale dispose de 1000 trafiquants qui partent tous le matin du jour 0 de la ville v_1. Quelle sera la proportion de trafiquants dans chacune des villes au bout d'une semaine ? d'un an ?

Méthode directe Il s'agit de calculer des puissances successives de $\mathbb{A}$ avec $\mathbf{x}^{(0)} = (1,0,0)^T$. Au bout d'une semaine on a

$$\mathbf{x}^{(7)} = \mathbb{A}^7 \cdot \mathbf{x}^{(0)} \simeq (56.4\%, 22.6\%, 21\%)^T.$$

Au but d'un an les proportions ne changent guère :

$$\mathbf{x}^{(365)} = \mathbb{A}^{365} \cdot \mathbf{x}^{(0)} \simeq (55.7\%, 22.9\%, 21.3\%)^T.$$

Le calcul de la puissance de $\mathbb{A}$ est lourd car il s'agit de 365 multiplications matricielles.

Méthode par diagonalisation Si on diagonalise $\mathbb{A}$ *i.e.* si on calcule $\mathbb{D}$ et $\mathbb{P}$ telles que $\mathbb{A} = \mathbb{P}\mathbb{D}\mathbb{P}^{-1}$ alors

$$\mathbf{x}^{(365)} = \mathbb{A}^{365} \cdot \mathbf{x}^{(0)} = \underbrace{(\mathbb{P}\mathbb{D}\mathbb{P}^{-1})(\mathbb{P}\mathbb{D}\mathbb{P}^{-1}) \cdots (\mathbb{P}\mathbb{D}\mathbb{P}^{-1})}_{365 \text{ fois}} \cdot \mathbf{x}^{(0)} = \mathbb{P}\mathbb{D}^{365}\mathbb{P}^{-1} \cdot \mathbf{x}^{(0)}.$$

Le calcul de la puissance de $\mathbb{D}$ est immédiat car

$$\mathbb{D} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1+2\sqrt{5}}{10} & 0 \\ 0 & 0 & \frac{1-2\sqrt{5}}{10} \end{pmatrix} \Longrightarrow \mathbb{D}^{365} = \begin{pmatrix} 1^{365} & 0 & 0 \\ 0 & \left(\frac{1+2\sqrt{5}}{10}\right)^{365} & 0 \\ 0 & 0 & \left(\frac{1-2\sqrt{5}}{10}\right)^{365} \end{pmatrix}$$

```
clc; clear all;
A = [0.8 0.3 0.2; 0.1 0.2 0.6; 0.1 0.5 0.2]
x_init=[1;0;0]
n=365

% Methode directe
x_365=A^n*x_init

% Methode par diagonalisation
[V,D] = eig(A);
x_365=V*D^n*inv(V)*x_init
```

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 $\mathbb{A}$, $r = \min\{n, p\}$, $\sigma_1 \geq \sigma_2 \geq \sigma_r \geq 0$. 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

$$\mathbb{A} = \mathbb{U}\mathbb{S}\mathbb{V}^T = \underbrace{\begin{pmatrix}\mathbf{u}_1 & \dots & \mathbf{u}_r & \mathbf{u}_{r+1} & \dots & \mathbf{u}_n\end{pmatrix}}_{n\times n} \underbrace{\begin{pmatrix}\sigma_1 & & & & & \\ & \ddots & & & & \\ & & \sigma_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0\end{pmatrix}}_{n\times p} \underbrace{\begin{pmatrix}\mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_r^T \\ \mathbf{v}_{r+1}^T \\ \vdots \\ \mathbf{v}_p^T\end{pmatrix}}_{p\times p}$$

$$= \underbrace{\begin{pmatrix}\mathbf{u}_1 & \dots & \mathbf{u}_r\end{pmatrix}}_{n\times r} \underbrace{\begin{pmatrix}\sigma_1 & & \\ & \ddots & \\ & & \sigma_r\end{pmatrix}}_{r\times r} \underbrace{\begin{pmatrix}\mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_r^T\end{pmatrix}}_{r\times p} = \sum_{i=1}^{r} \sigma_i \underbrace{\mathbf{u}_i \times \mathbf{v}_i^T}_{r\times r}$$

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 $\mathbb{A}^T = \mathbb{V}\mathbb{S}\mathbb{U}^T = \mathbb{V}\mathbb{S}\mathbb{U}^{-1}$ ainsi, pour $i = 1, \dots, r$, en multipliant par $\mathbb{A}$ à gauche $\mathbb{A}^T\mathbf{u}_i = \sigma_i\mathbf{v}_i$ et en multipliant par $\mathbb{A}$ à droite $\mathbb{A}\mathbf{v}_i = \sigma_i\mathbf{u}_i$ on obtient

$$\begin{aligned}\mathbb{A}\mathbb{A}^T\mathbf{u}_i &= \sigma_i\mathbb{A}\mathbf{v}_i = \sigma_i^2\mathbf{u}_i, \qquad \text{pour } i = 1, \dots, r \\ \mathbb{A}^T\mathbb{A}\mathbf{v}_i &= \sigma_i\mathbb{A}^T\mathbf{u}_i = \sigma_i^2\mathbf{v}_i, \qquad \text{pour } i = 1, \dots, r\end{aligned}$$

ainsi les σ_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 σ_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.

- ⋆ Pour stocker la matrice $\mathbb{A}$ nous avons besoin de $n \times p$ valeurs.
- ⋆ Pour stocker la décomposition SVD nous avons besoin de $n \times r + r + r \times p = (n + p + 1)r > (n + p + 1)r$ valeurs donc à priori on ne fait pas d'économies de stockage. Cependant, s'il existe $s < r$ tel que $\sigma_s = \sigma_{s+1} = \dots = \sigma_r = 0$, alors nous n'avons plus besoin que de $n \times s + s + s \times p = (n + p + 1)s$ valeurs. Si $s < np/(n + p + 1)$ on fait des économies de stockage.

Idée de la compression : si nous approchons $\mathbb{A}$ en ne gardant que les premiers s termes de la somme (sachant que les derniers termes sont multipliés par des σ_i plus petits, voire nuls)

$$\widetilde{\mathbb{A}} = \sum_{i=1}^{s} \sigma_i \underbrace{\mathbf{u}_i \times \mathbf{v}_i^T}_{\in\mathbb{R}^{n\times p}}, \qquad \text{où } s < r$$

- ⋆ pour stocker la matrice $\widetilde{\mathbb{A}}$ nous avons toujours besoin de $n \times p$ valeurs,
- ⋆ pour stocker la décomposition SVD nous avons besoin de $n \times s + s + s \times p = (n + p + 1)s$ valeurs. Si $s < np/(n + p + 1)$ on fait des économies de stockage.

1.4.4. 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

$$|\lambda| \leq ||\mathbb{A}|| \qquad \forall \lambda \in \sigma(\mathbb{A})$$

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 :

$$\sigma(\mathbb{A}) \subset \mathscr{O} = \subset \{ z \in \mathbb{C} \,|\, |z| \leq ||\mathbb{A}|| \}.$$

Parmi les normes les plus utilisées nous avons

$$||\mathbb{A}||_1 = \max_{j=1\ldots n} \sum_{i=1}^{n} |a_{ij}| = ||\mathbb{A}^T||_\infty$$
$$||\mathbb{A}||_\infty = \max_{i=1\ldots n} \sum_{j=1}^{n} |a_{ij}| = ||\mathbb{A}^T||_1$$

Une autre estimation est donnée par les disques de GERSHGORIN. Les disques de GERSHGORIN $\mathscr{R}_i$ et $\mathscr{C}_j$ associés à la i-ème ligne et à la j-ème colonne sont respectivement définis par

$$\mathscr{R}_i = \left\{ z \in \mathbb{C} \;\middle|\; |z - a_{ii}| \leq \sum_{\substack{j=1,\\ j\neq i}}^{n} |a_{ij}| \right\}, \qquad \mathscr{C}_j = \left\{ z \in \mathbb{C} \;\middle|\; |z - a_{jj}| \leq \sum_{\substack{i=1,\\ i\neq j}}^{n} |a_{ij}| \right\}.$$

Les disques de GERSHGORIN peuvent servir à localiser les valeurs propres d'une matrice, comme le montre la proposition suivante

Proposition 1.25
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 :

$$\sigma(\mathbb{A}) \subset \underbrace{\left(\bigcup_{i=1}^{n} \mathscr{R}_i \right)}_{\mathscr{S}_{\mathscr{R}}} \cap \underbrace{\left(\bigcup_{j=1}^{n} \mathscr{C}_j \right)}_{\mathscr{S}_{\mathscr{C}}}.$$

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

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

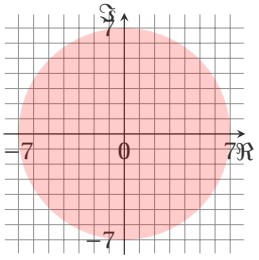
Nous avons les estimations suivantes :

1. Si on considère les normes $||\cdot||_1$ et $||\cdot||_\infty$ nous avons

$$||\mathbb{A}||_1 = \max_{j=1\ldots n} \sum_{i=1}^{n} |a_{ij}| = \max_{j=1\ldots n} \{\, |3| + |-1| + |0|; |2| + |2| + |1|; |3| + |-1| + |3| \,\} = \max_{j=1\ldots n} \{4;5;7\} = 7$$
$$||\mathbb{A}||_\infty = \max_{i=1\ldots n} \sum_{j=1}^{n} |a_{ij}| = \max_{i=1\ldots n} \{\, |3| + |2| + |3|; |-1| + |2| + |-1|; |0| + |1| + |3| \,\} = \max_{i=1\ldots n} \{8;4;4\} = 8$$

donc toutes les valeurs propres appartiennent au cercle de rayon 7 centré dans l'origine du plan complexe :

$$\sigma(\mathbb{A}) \subset \mathscr{O} = \{ z \in \mathbb{C} \,|\, |z| \leq 7 \}.$$



2. Disques des lignes :

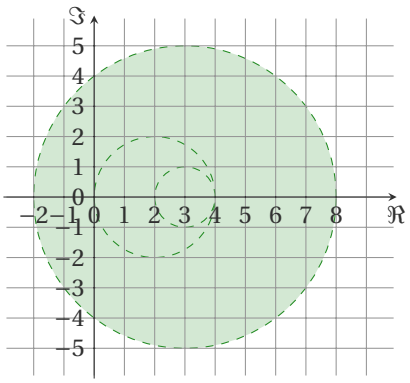
$$\mathscr{R}_1 = \left\{ z \in \mathbb{C} \;\middle|\; |z-a_{11}| \leq \sum_{\substack{j=1,\\ j\neq 1}}^{n} |a_{1j}| \right\} = \{ z \in \mathbb{C} \,|\, |z-3| \leq |2|+|3| \} = \{ z \in \mathbb{C} \,|\, |z-3| \leq 5 \},$$

$$\mathscr{R}_2 = \left\{ z \in \mathbb{C} \;\middle|\; |z-a_{22}| \leq \sum_{\substack{j=1,\\ j\neq 2}}^{n} |a_{2j}| \right\} = \{ z \in \mathbb{C} \,|\, |z-2| \leq |-1|+|-1| \} = \{ z \in \mathbb{C} \,|\, |z-2| \leq 2 \},$$

$$\mathscr{R}_3 = \left\{ z \in \mathbb{C} \;\middle|\; |z-a_{33}| \leq \sum_{\substack{j=1,\\ j\neq 3}}^{n} |a_{3j}| \right\} = \{ z \in \mathbb{C} \,|\, |z-3| \leq |0|+|1| \} = \{ z \in \mathbb{C} \,|\, |z-3| \leq 1 \}.$$

Toutes les valeurs propres appartiennent à la réunion des disques des lignes :

$$\sigma(\mathbb{A}) \subset \mathscr{S}_{\mathscr{R}} = \mathscr{R}_1 \cup \mathscr{R}_2 \cup \mathscr{R}_3 = \mathscr{R}_1.$$



3. Disques des colonnes :

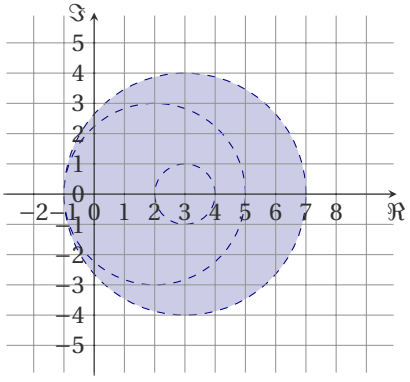
$$\mathscr{C}_1 = \left\{ z \in \mathbb{C} \;\middle|\; |z-a_{11}| \leq \sum_{\substack{i=1,\\ i\neq 1}}^{n} |a_{i1}| \right\} = \{ z \in \mathbb{C} \,|\, |z-3| \leq |-1|+|0| \} = \{ z \in \mathbb{C} \,|\, |z-3| \leq 1 \},$$

$$\mathscr{C}_2 = \left\{ z \in \mathbb{C} \;\middle|\; |z-a_{22}| \leq \sum_{\substack{i=1,\\ i\neq 2}}^{n} |a_{i2}| \right\} = \{ z \in \mathbb{C} \,|\, |z-2| \leq |2|+|1| \} = \{ z \in \mathbb{C} \,|\, |z-2| \leq 3 \},$$

$$\mathscr{C}_3 = \left\{ z \in \mathbb{C} \;\middle|\; |z-a_{33}| \leq \sum_{\substack{i=1,\\ i\neq 3}}^{n} |a_{i3}| \right\} = \{ z \in \mathbb{C} \,|\, |z-3| \leq |3|+|-1| \} = \{ z \in \mathbb{C} \,|\, |z-3| \leq 4 \}.$$

Toutes les valeurs propres appartiennent à la réunion des disques des colonnes :

$$\sigma(\mathbb{A}) \subset \mathscr{S}_{\mathscr{C}} = \mathscr{C}_1 \cup \mathscr{C}_2 \cup \mathscr{C}_3 = \mathscr{C}_3.$$



4. Toutes les valeurs propres appartiennent à l'intersection de ces trois régions :

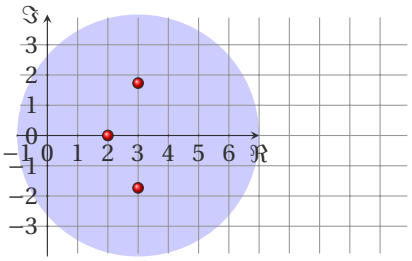
$$\sigma(\mathbb{A}) \subset \mathscr{S}_{\mathscr{O}} \cap \mathscr{S}_{\mathscr{R}} \cap \mathscr{S}_{\mathscr{C}} = \mathscr{C}_3$$

En effet, on a

$$p_{\mathbb{A}}(\lambda) = -\lambda^3+8\lambda^2-24\lambda+24 = -(\lambda-2)(\lambda^2-6\lambda+12)$$

donc

$$\lambda_1 = 2 \qquad \lambda_2 = 3+i\sqrt{3} \qquad \lambda_3 = \overline{\lambda_2} = 3-i\sqrt{3}.$$



EXEMPLE
Considérons la matrice

$$\mathbb{A} = \begin{pmatrix} 10 & 2 & 3 \\ -1 & 2 & -1 \\ 0 & 1 & 3 \end{pmatrix}.$$

Nous avons les estimations suivantes :

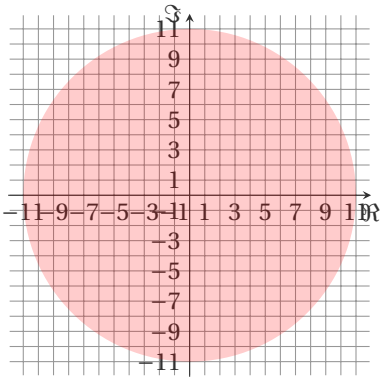
1. Si on considère les normes $||\cdot||_1$ et $||\cdot||_\infty$ nous avons

$$||\mathbb{A}||_1 = \max_{j=1\ldots n} \sum_{i=1}^{n} |a_{ij}| = \max_{j=1\ldots n} \{\,|10|+|-1|+|0|;|2|+|2|+|1|;|3|+|-1|+|3|\} = \max_{j=1\ldots n} \{11;5;7\} = 11$$

$$||\mathbb{A}||_\infty = \max_{i=1\ldots n} \sum_{j=1}^{n} |a_{ij}| = \max_{i=1\ldots n} \{\,|10|+|2|+|3|;|-1|+|2|+|-1|;|0|+|1|+|3|\} = \max_{i=1\ldots n} \{15;4;4\} = 15$$

donc toutes les valeurs propres appartiennent au cercle de rayon 11 centré dans l'origine du plan complexe :

$$\sigma(\mathbb{A}) \subset \mathscr{O} = \{z \in \mathbb{C} \,|\, |z| \leq 11\}.$$



2. Disques des lignes :

$$\mathscr{R}_1 = \left\{ z \in \mathbb{C} \;\middle|\; |z-a_{11}| \leq \sum_{\substack{j=1,\\ j\neq 1}}^{n} |a_{1j}| \right\} = \{z \in \mathbb{C} \,|\, |z-10| \leq |2|+|3|\} = \{z \in \mathbb{C} \,|\, |z-10| \leq 5\},$$

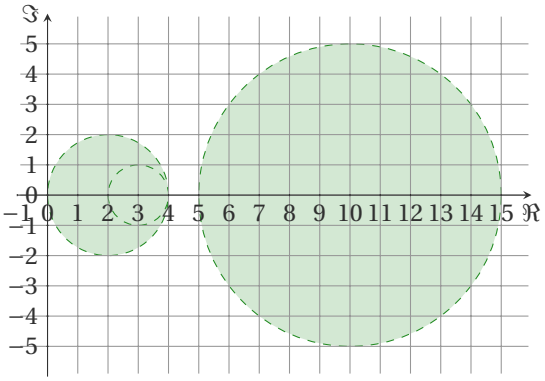
$$\mathscr{R}_2 = \left\{ z \in \mathbb{C} \;\middle|\; |z-a_{22}| \leq \sum_{\substack{j=1,\\ j\neq 2}}^{n} |a_{2j}| \right\} = \{z \in \mathbb{C} \,|\, |z-2| \leq |-1|+|-1|\} = \{z \in \mathbb{C} \,|\, |z-2| \leq 2\},$$

$$\mathscr{R}_3 = \left\{ z \in \mathbb{C} \;\middle|\; |z-a_{33}| \leq \sum_{\substack{j=1,\\ j\neq 3}}^{n} |a_{3j}| \right\} = \{z \in \mathbb{C} \,|\, |z-3| \leq |0|+|1|\} = \{z \in \mathbb{C} \,|\, |z-3| \leq 1\}.$$

Toutes les valeurs propres appartiennent à la réunion des disques des lignes :

$$\sigma(\mathbb{A}) \subset \mathscr{S}_{\mathscr{R}} = \mathscr{R}_1 \cup \mathscr{R}_2 \cup \mathscr{R}_3.$$

De plus, comme le disque $\mathscr{R}_1$ est disjoint de la réunion $\mathscr{R}_2 \cup \mathscr{R}_3$, une et une seule valeur propre est contenue dans $\mathscr{R}_1$ et les deux autres valeurs propres appartiennent à $\mathscr{R}_2 \cup \mathscr{R}_3 = \mathscr{R}_2$.



3. Disques des colonnes :

$$\mathscr{C}_1 = \left\{ z \in \mathbb{C} \,\middle|\, |z - a_{11}| \leq \sum_{\substack{i=1,\\ i\neq 1}}^{n} |a_{i1}| \right\} = \{ z \in \mathbb{C} \,|\, |z-10| \leq |-1| + |0| \} = \{ z \in \mathbb{C} \,|\, |z-10| \leq 1 \},$$

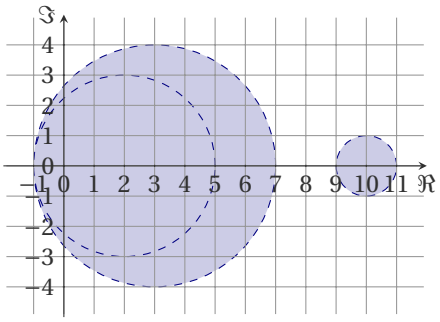
$$\mathscr{C}_2 = \left\{ z \in \mathbb{C} \,\middle|\, |z - a_{22}| \leq \sum_{\substack{i=1,\\ i\neq 2}}^{n} |a_{i2}| \right\} = \{ z \in \mathbb{C} \,|\, |z-2| \leq |2| + |1| \} = \{ z \in \mathbb{C} \,|\, |z-2| \leq 3 \},$$

$$\mathscr{C}_3 = \left\{ z \in \mathbb{C} \,\middle|\, |z - a_{33}| \leq \sum_{\substack{i=1,\\ i\neq 3}}^{n} |a_{i3}| \right\} = \{ z \in \mathbb{C} \,|\, |z-3| \leq |3| + |-1| \} = \{ z \in \mathbb{C} \,|\, |z-3| \leq 4 \}.$$

Toutes les valeurs propres appartiennent à la réunion des disques des colonnes :

$$\sigma(\mathbb{A}) \subset \mathscr{S}_{\mathscr{C}} = \mathscr{C}_1 \cup \mathscr{C}_2 \cup \mathscr{C}_3.$$

De plus, comme le disque $\mathscr{C}_1$ est disjoint de la réunion $\mathscr{C}_2 \cup \mathscr{C}_3$, une et une seule valeur propre est contenue dans $\mathscr{C}_1$ et les deux autres valeurs propres appartiennent à $\mathscr{C}_2 \cup \mathscr{C}_3 = \mathscr{C}_3$.



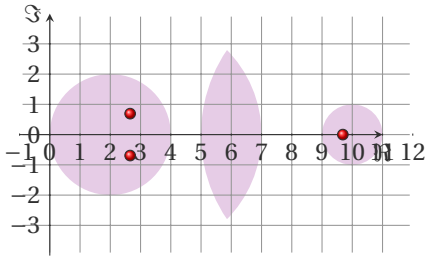
4. Toutes les valeurs propres appartiennent à l'intersection de ces trois régions :

$$\sigma(\mathbb{A}) \subset \mathscr{S}_{\mathscr{C}} \cap \mathscr{S}_{\mathscr{R}} \cap \mathscr{S}_{\mathscr{C}}$$

En effet, on a

$$\lambda_1 \simeq 9.6876 \qquad \lambda_2 \simeq 2.6562 + 0.6928i \qquad \lambda_3 = \overline{\lambda_2} \simeq 2.6562 - 0.6928i.$$

```
A=[10 2 3; -1 2 -1; 0 1 3]
eig(A)
```



1.5. Exercices

1.5.1. Calcul matriciel

Exercice 1.1 (Écriture matricielle)
On considère les matrices $\mathbb{A} = (a_{ij})$, $\mathbb{B} = (b_{ij})$ et $\mathbb{C} = (c_{ij})$ carrées d'ordre 4 définies par $a_{ij} = i^2$, $b_{ij} = i + j$, $c_{ij} = \min\{i; j\}$. Écrire ces matrices sous la forme de tableaux de nombres.

Correction

$$\mathbb{A} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 4 & 4 & 4 & 4 \\ 9 & 9 & 9 & 9 \\ 16 & 16 & 16 & 16 \end{pmatrix} \qquad \mathbb{B} = \begin{pmatrix} 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 7 \\ 5 & 6 & 7 & 8 \end{pmatrix} \qquad \mathbb{C} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 \\ 1 & 2 & 3 & 4 \end{pmatrix}$$

```
A=zeros(4);
for i=1:4
   A(i,:)=i^2;
end
A
```

```
B=zeros(4);
for i=1:4
   for j=1:4
      B(i,j)=i+j;
   end
end
B
```

```
C=zeros(4);
for i=1:4
   for j=1:4
      C(i,j)=min(i,j);
   end
end
C
```

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 $\mathbb{A} - 2\mathbb{B} - \mathbb{C} = \mathbb{0}$.
2. Trouver une matrice $\mathbb{D}$ telle que $\mathbb{A} + \mathbb{B} + \mathbb{C} - 4\mathbb{D} = \mathbb{0}$.

Correction

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

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

   $\mathbb{A} - 2\mathbb{B} - \mathbb{C} = \mathbb{0}$.

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

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

Correction

$$\overbrace{\begin{pmatrix}3&1&5\\2&7&0\end{pmatrix}}^{2\times3}\overbrace{\begin{pmatrix}2&1&-1&0\\3&0&1&8\\0&-5&3&4\end{pmatrix}}^{3\times4}=\overbrace{\begin{pmatrix}3\times2+1\times3+5\times0&3\times1+1\times0+5\times(-5)&3\times(-1)+1\times1+5\times3&3\times0+1\times8+5\times4\\2\times2+7\times3+0\times0&2\times1+7\times0+0\times(-5)&2\times(-1)+7\times1+0\times3&2\times0+7\times8+0\times4\end{pmatrix}}^{2\times4}$$

$$=\begin{pmatrix}9&-220&13&28\\25&2&5&560\end{pmatrix}$$

$$\overbrace{\begin{pmatrix}-3&0&5\end{pmatrix}}^{1\times3}\overbrace{\begin{pmatrix}2\\-4\\-3\end{pmatrix}}^{3\times1}=\overbrace{\begin{pmatrix}-3\times2+0\times(-4)+5\times(-3)\end{pmatrix}}^{1\times1}=-21$$

$$\overbrace{\begin{pmatrix}-3\\0\\5\end{pmatrix}}^{3\times1}\overbrace{\begin{pmatrix}2&-4&-3\end{pmatrix}}^{1\times3}=\overbrace{\begin{pmatrix}-3\times2&-3\times(-4)&-3\times(-3)\\0\times2&0\times(-4)&0\times(-3)\\5\times2&5\times(-4)&5\times(-3)\end{pmatrix}}^{3\times3}=\begin{pmatrix}-6&12&9\\0&0&0\\10&-20&-15\end{pmatrix}$$

```
[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]
[-3 0 5]'*[2 -4 -3] % ce qui equivaut a [-3; 0; 5]*[2 -4 -3]
```

Exercice 1.4
Soit les matrices

$$\mathbb{A}=\begin{pmatrix}1&2&3\\-1&0&1\\0&1&0\end{pmatrix}\qquad \mathbb{B}=\begin{pmatrix}2&-1&0\\-1&0&1\end{pmatrix}\qquad \mathbf{u}=\begin{pmatrix}1\\x\\x^2\end{pmatrix}\qquad \mathbf{v}=\begin{pmatrix}1\\0\\1\end{pmatrix}$$

1. Calculer tous les produits possibles à partir de $\mathbb{A}$, $\mathbb{B}$, $\mathbf{u}$ et $\mathbf{v}$.
2. Calculer $(\mathbb{A}-\mathbb{I})^7$ et en extraire le coefficient en position $(2,3)$.
3. Calculer $\mathbb{A}^{-1}$ et la trace de $\mathbb{A}^{-1}$ (*i.e.* la somme des coefficients sur la diagonale).

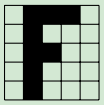
Correction
Sans utiliser un module spécifique, il n'est pas possible de faire des calculs formels avec MATLAB/Octave, donc on ne peut pas utiliser $\mathbf{u}$ sans donner une valeur numérique à x.

```
A=[1 2 3; -1 0 1; 0 1 0] % 3x3
B=[2 -1 0; -1 0 1] % 2x3
v=[1;0;1] % 3x1
% Les produits possibles sont
B*A % (2x3)*(3x3) -> 2x3
A*v % (3x3)*(3x1) -> 3x1
```

```
B*v % (2x3)*(3x1) -> 2x1
%
Id=eye(3)
D=(A-Id)^7 % c'est bien ^7 (produit matriciel) et non .^7
D(2,3) % -> 153
%
invA=A^(-1) % ou inv(A)
sum(diag(invA))
```

Exercice 1.5 (Multiplication matricielle appliquée)

On modélise une image en noir et blanc formée de 25 pixels par une matrice de 5 lignes et 5 colonnes, dans laquelle 0 correspond à un pixel blanc et 1 à un pixel noir. L'image à modéliser est la suivante :

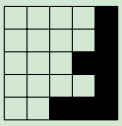


1. Donner la matrice $\mathbb{M}$ associée à l'image.
2. Soient

$$\mathbb{A} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix} \qquad \mathbb{B} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

   Calculer le produit $\mathbb{AM}$. Quel est l'effet de la matrice $\mathbb{A}$ sur l'image ? Quel est l'effet si on fait le produit $\mathbb{MA}$ sur l'image ?

   Calculer le produit $\mathbb{BM}$. Quel est l'effet de la matrice $\mathbb{B}$ sur l'image ? Quel est l'effet si on fait le produit $\mathbb{MB}$ sur l'image ?
3. Quelle image obtient-on en faisant le produit $\mathbb{AMB}$?
4. Quel produit matriciel peut-on faire pour obtenir la figure suivante ?



Correction

1.

$$\mathbb{M} = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

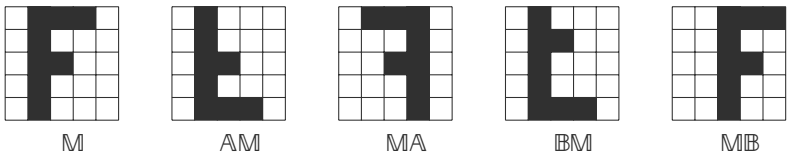
2. Le produit $\mathbb{AM}$ correspond à inverser l'ordre des lignes de la matrice $\mathbb{M}$: l'image est alors symétrique par rapport à la troisième ligne.

   Le produit $\mathbb{MA}$ correspond à inverser l'ordre des colonnes de la matrice $\mathbb{M}$: l'image est alors symétrique par rapport à la troisième colonne.

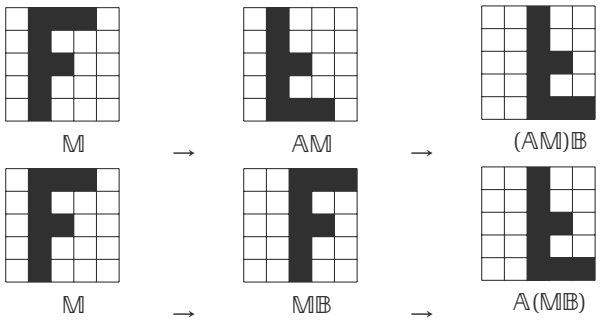
   Le produit $\mathbb{BM}$ correspond à translater les lignes de la matrice $\mathbb{M}$ d'un rang vers le haut (la première ligne passant en cinquième ligne) : l'image est alors translatée d'un rang vers le haut (la première ligne passant en cinquième ligne).

   Le produit $\mathbb{MB}$ correspond à translater les colonnes de la matrice $\mathbb{M}$ d'un rang vers la droite (la cinquième colonne passant en première colonne) : l'image est alors translatée d'un rang vers la droite (la cinquième colonne passant en première colonne).

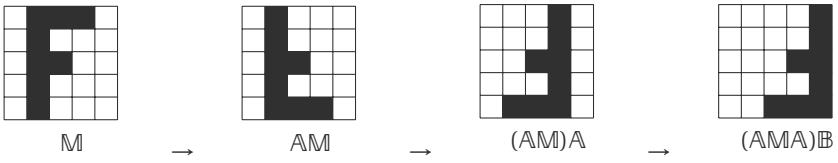
$$\mathbb{AM} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \end{pmatrix} \quad \mathbb{MA} = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \quad \mathbb{BM} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \end{pmatrix} \quad \mathbb{MB} = \begin{pmatrix} 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$



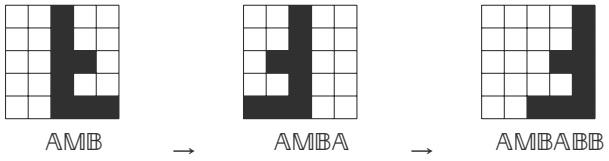
3. Le produit AMB = (AM)B correspond par exemple à une symétrie horizontale par rapport à la troisième ligne suivie d'une translation d'un rang vers la droite. On peut aussi l'écrire comme AMB = A(MB) qui correspond à une translation d'un rang vers la droite suivie d'une symétrie horizontale par rapport à la troisième ligne. Dans tous les cas on obtient l'image suivante :



4. On peut par exemple calculer AMAB.



Ou encore calculer AMBABB.




```
M=[0 1 1 1 0
   0 1 0 0 0
   0 1 1 0 0
   0 1 0 0 0
   0 1 0 0 0]
```

```
A=eye(5);
A=A(:,5:-1:1)
%
B=diag(ones(4,1),1);
B(5,1)=1;
B
```

```
A*M
M*A
B*M
M*B
A*M*B
A*M*A*B
A*M*B*A*B*B
```

Exercice 1.6
Calculer a, b, c et d tels que

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

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 1.7

On dit que deux matrices $\mathbb{A}$ et $\mathbb{B}$ commutent si $\mathbb{AB} = \mathbb{BA}$. 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

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

Si de plus on veut que $\mathbb{AB} = \mathbb{BA} = \mathbb{I}_3$, *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 1.8
Trouver pour quelles valeurs de $t \in \mathbb{R}$ les matrices suivantes sont inversibles :

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

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 $t \in \mathbb{R}\setminus\{-3,3\}$.

```
determinant=@(t)[det([t+3, t^2-9; t^2+9, t-3])];
fsolve(determinant,1)
fsolve(determinant,-1)
```

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

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

```
determinant=@(t)[det([t^2-9, t+3; t-3, t^2+9])];
fsolve(determinant,1)
fsolve(determinant,-1)
```

Exercice 1.9
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 :

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

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

```
determinant=@(t)[det([t+3, -1, 1; 5, t-3, 1; 6, -6, t+4])];
fsolve(determinant,1)
fsolve(determinant,-1)
fsolve(determinant,-10)
```

Exercice 1.10
Soit a, b et c trois réels quelconques, calculer les déterminants suivants :

$$D_1 = \det\begin{pmatrix} 1 & 1 & 1 \\ a & b & c \\ a^2 & b^2 & c^2 \end{pmatrix} \qquad D_2 = \det\begin{pmatrix} 1+a & 1 & 1 \\ 1 & 1+a & 1 \\ 1 & 1 & 1+a \end{pmatrix}$$

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

$$D_1 = \det\begin{pmatrix} 1 & 1 & 1 \\ a & b & c \\ a^2 & b^2 & c^2 \end{pmatrix} \xrightarrow[=]{\substack{C_2\leftarrow C_2-C_1 \\ C_3\leftarrow C_3-C_1}} \det\begin{pmatrix} 1 & 0 & 0 \\ a & b-a & c-a \\ a^2 & b^2-a^2 & c^2-a^2 \end{pmatrix} = \det\begin{pmatrix} b-a & c-a \\ b^2-a^2 & c^2-a^2 \end{pmatrix}$$
$$= (b-a)(c^2-a^2) - (c-a)(b^2-a^2) = (b-a)(c-a)\Big((c+a)-(b+a)\Big) = (b-a)(c-a)(c-b);$$
$$D_2 = \det\begin{pmatrix} 1+a & 1 & 1 \\ 1 & 1+a & 1 \\ 1 & 1 & 1+a \end{pmatrix} \xrightarrow[=]{\substack{C_2\leftarrow C_2-C_1 \\ C_3\leftarrow C_3-C_1}} \det\begin{pmatrix} 1+a & -a & -a \\ 1 & a & 0 \\ 1 & 0 & a \end{pmatrix} \xrightarrow[=]{L_1\leftarrow L_1+L_2+L_3} \det\begin{pmatrix} 3+a & 0 & 0 \\ 1 & a & 0 \\ 1 & 0 & a \end{pmatrix} = a^2(3+a).$$

Exercice 1.11

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 des matrices $\mathbb{B} = \begin{pmatrix} 1 & 2 & 8 \\ 2 & 1 & 4 \\ 0 & 3 & 12 \end{pmatrix}$ et $\mathbb{C} = \begin{pmatrix} 2 & 1 & 3 \\ 8 & 4 & 12 \\ 1 & 2 & 0 \end{pmatrix}$.
3. Calculer le déterminant des matrices $\mathbb{D} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 2 & 3 & 7 & 4 \\ 3 & 1 & 12 & 0 \\ 4 & 0 & -5 & 0 \end{pmatrix}$ et $\mathbb{E} = \begin{pmatrix} 0 & 2 & 3 & 4 \\ 1 & 7 & 12 & -5 \\ 0 & 3 & 1 & 0 \\ 0 & 4 & 0 & 0 \end{pmatrix}$.

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 $1 \leq \mathrm{rg}(\mathbb{B}) \leq 3$. Comme $\det(\mathbb{B}) = 0$ (sans faire de calcul, il suffit de remarquer que $C_3 = 4C_2$), alors $1 \leq \mathrm{rg}(\mathbb{B}) \leq 2$. Comme $\det\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} = -3 \neq 0$, on conclut que $\mathrm{rg}(\mathbb{B}) = 2$. De la même manière, $1 \leq \mathrm{rg}(\mathbb{C}) \leq 3$ et puisque $\det(\mathbb{C}) = 0$ (sans faire de calcul, il suffit de remarquer que $L_2 = 4L_1$), alors $1 \leq \mathrm{rg}(\mathbb{C}) \leq 2$. Comme $\det\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} = 3 \neq 0$, on conclut que $\mathrm{rg}(\mathbb{C}) = 2$.

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

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

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

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

Exercice 1.12 (F. Le Roux)

En admettant le fait que les nombres 2001, 1073, 5800 et 8903 sont tous divisibles par 29, montrer que le déterminant de la matrice

$$\mathbb{A} = \begin{pmatrix} 2 & 0 & 0 & 1 \\ 1 & 0 & 7 & 3 \\ 5 & 8 & 0 & 0 \\ 8 & 9 & 0 & 3 \end{pmatrix}$$

est aussi divisible par 29 (*sans* calculer ce déterminant!).

Correction
Le déterminant ne change pas lorsque on ajoute à une colonne une combinaison linéaire des autres colonnes. Si on ajoute à la quatrième colonne la combinaison linéaire $\sum_{i=1}^{3} 10^{4-i} C_i$ on obtient

$$\sum_{i=1}^{3} 10^{4-i} C_i + C_4 = 10^3 \begin{pmatrix} 2 \\ 1 \\ 5 \\ 8 \end{pmatrix} + 10^2 \begin{pmatrix} 0 \\ 0 \\ 8 \\ 9 \end{pmatrix} + 10 \begin{pmatrix} 0 \\ 7 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 1 \\ 3 \\ 0 \\ 3 \end{pmatrix} = \begin{pmatrix} 2001 \\ 1073 \\ 5800 \\ 8903 \end{pmatrix} = 29 \begin{pmatrix} 69 \\ 37 \\ 200 \\ 307 \end{pmatrix}$$

donc

$$\begin{aligned} \det(\mathbb{A}) &= \det \begin{pmatrix} 2 & 0 & 0 & 29\times 69 \\ 1 & 0 & 7 & 29\times 37 \\ 5 & 8 & 0 & 29\times 200 \\ 8 & 9 & 0 & 29\times 307 \end{pmatrix} \\ &= -29\times 69\times \det \begin{pmatrix} 1 & 0 & 7 \\ 5 & 8 & 0 \\ 8 & 9 & 0 \end{pmatrix} + 29\times 37\times \det \begin{pmatrix} 2 & 0 & 0 \\ 5 & 8 & 0 \\ 8 & 9 & 0 \end{pmatrix} - 29\times 200\times \det \begin{pmatrix} 2 & 0 & 0 \\ 1 & 0 & 7 \\ 8 & 9 & 0 \end{pmatrix} + 29\times 307\times \det \begin{pmatrix} 2 & 0 & 0 \\ 1 & 0 & 7 \\ 5 & 8 & 0 \end{pmatrix} \\ &= 29\times \left[-69\times \det \begin{pmatrix} 1 & 0 & 7 \\ 5 & 8 & 0 \\ 8 & 9 & 0 \end{pmatrix} + 37\times \det \begin{pmatrix} 2 & 0 & 0 \\ 5 & 8 & 0 \\ 8 & 9 & 0 \end{pmatrix} - 200\times \det \begin{pmatrix} 2 & 0 & 0 \\ 1 & 0 & 7 \\ 8 & 9 & 0 \end{pmatrix} + 307\times \det \begin{pmatrix} 2 & 0 & 0 \\ 1 & 0 & 7 \\ 5 & 8 & 0 \end{pmatrix} \right] \end{aligned}$$

Exercice 1.13
Soit $n \geq 2$ un entier naturel pair. Une matrice de taille $n \times n$ est à remplir par deux joueurs, A (qui veut un déterminant non nul, commence) et B (qui veut un déterminant nul). Ils ont le droit de mettre n'importe quel réel dans une case vide de la matrice, chacun leur tour. Trouver une stratégie gagnante pour B.

Correction
Il suffit de faire en sorte qu'une colonne soit identique (ou un multiple) d'une autre.

1.5.2. Espaces vectoriels et bases

Exercice 1.14
Démontrer que l'ensemble

$$F = \left\{ (x,y,z) \in \mathbb{R}^3 \,\middle|\, x+y+2z = 0 \right\}$$

est un sous-espace vectoriel de $\mathbb{R}^3$.

Correction
On montre que $F = \text{Vect}\{\mathbf{e}_1, \ldots, \mathbf{e}_p\}$ où $\mathbf{e}_1, \ldots, \mathbf{e}_p$ sont des éléments de E. En effet

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \in F \quad \Longleftrightarrow \quad \begin{cases} x,y,z \in \mathbb{R} \\ x+y+2z = 0 \end{cases} \quad \Longleftrightarrow \quad \begin{cases} x \in \mathbb{R} \\ z \in \mathbb{R} \\ y = -2z - x. \end{cases}$$

Donc

$$F = \left\{ \begin{pmatrix} \kappa_1 \\ -\kappa_1 - 2\kappa_2 \\ \kappa_2 \end{pmatrix} \,\middle|\, \kappa_1, \kappa_2 \in \mathbb{R} \right\} = \left\{ \kappa_1 \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} + \kappa_2 \begin{pmatrix} 0 \\ -2 \\ 1 \end{pmatrix} \right\} = \text{Vect} \left\{ \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ -2 \\ 1 \end{pmatrix} \right\}.$$

Par conséquent F est un sous-espace vectoriel de $\mathbb{R}^3$. (Cette méthode permet également d'en déduire que $\{(1,-1,0),(0,-2,1)\}$ est une famille génératrice de F.)

Exercice 1.15
Démontrer que l'ensemble

$$F = \left\{ \begin{pmatrix} a & b \\ b & c \end{pmatrix} \in \mathscr{M}_2(\mathbb{R}) \;\middle|\; a+b=0 \right\}$$

est un sous-espace vectoriel de $\mathscr{M}_2(\mathbb{R})$.

Correction

$$F = \left\{ \begin{pmatrix} a & b \\ b & c \end{pmatrix} \in \mathscr{M}_2(\mathbb{R}) \;\middle|\; a+b=0 \right\} = \left\{ a\begin{pmatrix} 1 & -1 \\ -1 & 0 \end{pmatrix} + c\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \;\middle|\; a,c \in \mathbb{R} \right\} = \operatorname{Vect}\left\{ \begin{pmatrix} 1 & -1 \\ -1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \right\}.$$

Par conséquent F est un sous-espace vectoriel de $\mathscr{M}_2(\mathbb{R})$.

Exercice 1.16
Démontrer que l'ensemble

$$F = \left\{ (x,x,y) \in \mathbb{R}^3 \;\middle|\; x,y \in \mathbb{R} \right\}$$

est un sous-espace vectoriel de $\mathbb{R}^3$.

Correction

$$F = \left\{ \begin{pmatrix} \kappa_1 \\ \kappa_1 \\ \kappa_2 \end{pmatrix} \;\middle|\; \kappa_1, \kappa_2 \in \mathbb{R} \right\} = \left\{ \kappa_1 \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + \kappa_2 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \;\middle|\; \kappa_1, \kappa_2 \in \mathbb{R} \right\} = \operatorname{Vect}\left\{ \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}.$$

Par conséquent F est un sous-espace vectoriel de $\mathbb{R}^3$.

Exercice 1.17
Démontrer que l'ensemble

$$F = \left\{ \begin{pmatrix} a & b \\ b & b \end{pmatrix} \;\middle|\; a,b \in \mathbb{R} \right\}$$

est un sous-espace vectoriel de $\mathscr{M}_2(\mathbb{R})$.

Correction

$$F = \left\{ \begin{pmatrix} a & b \\ b & b \end{pmatrix} \;\middle|\; a,b \in \mathbb{R} \right\} = \left\{ a\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + b\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \;\middle|\; a,b \in \mathbb{R} \right\} = \operatorname{Vect}\left\{ \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \right\}.$$

Par conséquent F est un sous-espace vectoriel de $\mathscr{M}_2(\mathbb{R})$.

Exercice 1.18
Démontrer que l'ensemble

$$F = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathscr{M}_2(\mathbb{R}) \;\middle|\; a+b+c+d=0 \right\}$$

est un sous-espace vectoriel de $\mathscr{M}_2(\mathbb{R})$.

Correction

$$\begin{aligned} F &= \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathscr{M}_2(\mathbb{R}) \;\middle|\; a+b+c+d=0 \right\} = \left\{ \begin{pmatrix} a & b \\ c & -a-b-c \end{pmatrix} \;\middle|\; a,b,c \in \mathbb{R} \right\} \\ &= \left\{ a\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} + b\begin{pmatrix} 0 & 1 \\ 0 & -1 \end{pmatrix} + c\begin{pmatrix} 0 & 0 \\ 1 & -1 \end{pmatrix} \;\middle|\; a,b,c \in \mathbb{R} \right\} = \operatorname{Vect}\left\{ \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 0 & -1 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 1 & -1 \end{pmatrix} \right\}. \end{aligned}$$

Par conséquent F est un sous-espace vectoriel de $\mathscr{M}_2(\mathbb{R})$.

Exercice 1.19
Démontrer que l'ensemble

$$F = \left\{ a + bx + cx^2 \in \mathbb{R}_2[x] \;\middle|\; a + b + 2c = 0 \right\}$$

est un sous-espace vectoriel de $\mathbb{R}_2[x]$.

Correction
On montre que $F = \text{Vect}\left\{\mathbf{e}_1, \ldots, \mathbf{e}_p\right\}$ où $\mathbf{e}_1, \ldots, \mathbf{e}_p$ sont des éléments de $\mathbb{R}_2[x]$. En effet

$$\begin{aligned} F &= \left\{ a + bx + cx^2 \in \mathbb{R}_2[x] \;\middle|\; a + b + 2c = 0 \right\} \\ &= \left\{ a + (-2c - a)x + cx^2 \;\middle|\; a, c \in \mathbb{R} \right\} \\ &= \left\{ a(1 - x) + c(-2x + x^2) \;\middle|\; a, c \in \mathbb{R} \right\} \\ &= \text{Vect}\left\{ 1 - x, -2x + x^2 \right\}. \end{aligned}$$

Par conséquent F est un sous-espace vectoriel de $\mathbb{R}_2[x]$.
(On peut également en déduire que $\{1 - x, -2x + x^2\}$ est une famille génératrice de F.)

Exercice 1.20
Démontrer que l'ensemble

$$F = \left\{ p \in \mathbb{R}_2[x] \;\middle|\; p(1) = 0 \right\}$$

est un sous-espace vectoriel de $\mathbb{R}_2[x]$.

Correction
On montre que $F = \text{Vect}\left\{\mathbf{e}_1, \ldots, \mathbf{e}_p\right\}$ où $\mathbf{e}_1, \ldots, \mathbf{e}_p$ sont des éléments de $\mathbb{R}_2[x]$. En effet

$$\begin{aligned} F &= \left\{ a + bx + cx^2 \in \mathbb{R}_2[x] \;\middle|\; a + b + c = 0 \right\} \\ &= \left\{ a + bx + (-a - b)x^2 \;\middle|\; a, c \in \mathbb{R} \right\} \\ &= \left\{ a(1 - x^2) + b(x - x^2) \;\middle|\; a, b \in \mathbb{R} \right\} \\ &= \text{Vect}\left\{ 1 - x^2, x - x^2 \right\}. \end{aligned}$$

Par conséquent F est un sous-espace vectoriel de $\mathbb{R}_2[x]$.
(On peut également en déduire que $\left\{1 - x^2, x - x^2\right\}$ est une famille génératrice de F.)

Exercice 1.21
Démontrer que l'ensemble

$$F = \left\{ p \in \mathbb{R}_2[x] \;\middle|\; p'(1) = 0 \right\}$$

est un sous-espace vectoriel de $\mathbb{R}_2[x]$.

Correction
On montre que $F = \text{Vect}\left\{\mathbf{e}_1, \ldots, \mathbf{e}_p\right\}$ où $\mathbf{e}_1, \ldots, \mathbf{e}_p$ sont des éléments de $\mathbb{R}_2[x]$. En effet

$$\begin{aligned} F &= \left\{ a + bx + cx^2 \in \mathbb{R}_2[x] \;\middle|\; b + 2c = 0 \right\} \\ &= \left\{ a - 2cx + cx^2 \;\middle|\; a, c \in \mathbb{R} \right\} \\ &= \left\{ a + c(-2x + x^2) \;\middle|\; a, c \in \mathbb{R} \right\} \\ &= \text{Vect}\left\{ 1, -2x + x^2 \right\}. \end{aligned}$$

Par conséquent F est un sous-espace vectoriel de $\mathbb{R}_2[x]$.
(On peut également en déduire que $\left\{1, -2x + x^2\right\}$ est une famille génératrice de F.)

Exercice 1.22
Montrer que l'ensemble

$$F = \left\{ \mathbb{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathscr{M}_2(\mathbb{R}) \;\middle|\; \det(\mathbb{A}) = 0 \right\}$$

n'est pas un sous-espace vectoriel de $\mathscr{M}_2(\mathbb{R})$.

Correction
Soit $\mathbb{A} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ et $\mathbb{A}' = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ deux matrices de F. Comme $\mathbb{A} + \mathbb{A}' = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$, alors $\det(\mathbb{A} + \mathbb{A}') = 1$, donc $\mathbb{A} + \mathbb{A}' \notin F$.

Exercice 1.23
Prouver que les familles suivantes sont libres :

1. $\mathscr{A} = \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\} \subset \mathbb{R}_2$
2. $\mathscr{B} = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix} \right\} \subset \mathscr{M}_3(\mathbb{R})$
3. $\mathscr{C} = \left\{ 1, t, t^2 \right\} \subset \mathbb{R}_2[t]$
4. $\mathscr{D} = \{ 1, t, t(t-1), t(t-1)(t-2) \} \subset \mathbb{R}_3[t]$

Correction

1. $\alpha \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \beta \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ ssi $\begin{pmatrix} \alpha+\beta \\ \beta \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ ssi $\alpha = \beta = 0$ donc la famille est libre.
2. $\alpha \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \beta \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} + \gamma \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} + \delta \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \implies \begin{cases} \alpha = 0 \\ \beta + \gamma = 0 \\ \gamma + \delta = 0 \\ \delta = 0 \end{cases} \implies \alpha = \beta = \gamma = \delta = 0$ donc la famille est libre.
3. C'est la base canonique de $\mathbb{R}_3[t]$ donc la famille est libre.
4. $\alpha + \beta t + \gamma t(t-1) + \delta t(t-1)(t-2) = 0$ pour tout $t \in \mathbb{R}$ ssi $\alpha + (\beta - \gamma + 2\delta)t + (\gamma - 3\delta)t^2 + \delta t^3 = 0$ pour tout $t \in \mathbb{R}$ ssi $\alpha = \beta = \gamma = \delta = 0$ donc la famille est libre.

Exercice 1.24
Montrer que l'ensemble

$$F = \left\{ \begin{pmatrix} a & b \\ b & c \end{pmatrix} \in \mathscr{M}_2(\mathbb{R}) \;\middle|\; ac = b^2 \right\}$$

n'est pas un sous-espace vectoriel de $\mathscr{M}_2(\mathbb{R})$.

Correction
Soit $\mathbb{A} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ et $\mathbb{A}' = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ deux matrices de F. Comme $\mathbb{A} + \mathbb{A}' = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$, alors $\det(\mathbb{A} + \mathbb{A}') = 1$, donc $\mathbb{A} + \mathbb{A}' \notin F$.

Exercice 1.25
Montrer que l'ensemble

$$F = \left\{ (x, y, z, w) \in \mathbb{R}^4 \;\middle|\; ax + by - z = 0,\ bx + y - w \geq 0 \ \forall a, b \in \mathbb{R} \right\}$$

n'est pas un espace vectoriel.

Correction
Le vecteur $\mathbf{v} = (0, 0, 0, -1) \in F$ mais $-\mathbf{v} \notin F$ donc F n'est pas un espace vectoriel.

Exercice 1.26
Montrer que l'ensemble

$$F = \left\{ (x, y, z) \in \mathbb{R}^3 \;\middle|\; x^2 + y = 0 \right\}$$

n'est pas un espace vectoriel.

Correction
Soit $\mathbf{u} = (a, b, c)$ et $\mathbf{v} = (d, e, f)$ deux vecteurs de l'ensemble F. Alors $a^2 + b = 0$ et $d^2 + e = 0$. Pour que la somme $\mathbf{u} + \mathbf{v} = (a + d, b + e, c + f)$ appartient à F il faut vérifier si $(a + d)^2 + (b + e) = 0$. Or on a $a^2 + d^2 + 2ad + b + e = 2ad$ qui n'est pas forcement 0 donc F n'est pas un espace vectoriel.

Exercice 1.27
Soient $\mathbf{u}_1 = (1,1,1)$ et $\mathbf{u}_2 = (1,2,3)$ deux vecteurs de $\mathbb{R}^3$. Trouver une condition nécessaire et suffisante sur les réels x, y, z pour que le vecteur $\mathbf{w} = (x,y,z)$ appartient à $\text{Vect}\{\mathbf{u}_1, \mathbf{u}_2\}$.

Correction

$$\begin{aligned}
\mathbf{w} \in \text{Vect}\{\mathbf{u}_1,\mathbf{u}_2\} &\iff \exists\,(a,b)\in\mathbb{R}^2 \text{ tel que } \mathbf{w} = a\mathbf{u}_1 + b\mathbf{u}_2 \\
&\iff \exists\,(a,b)\in\mathbb{R}^2 \text{ tel que } \begin{cases} a+b=x, \\ a+2b=y, \\ a+3b=z \end{cases} \\
&\iff \exists\,(a,b)\in\mathbb{R}^2 \text{ tel que } \begin{cases} a=2x-y, \\ b=y-x, \\ z=-x+2y \end{cases} \\
&\iff x-2y+z=0.
\end{aligned}$$

Exercice 1.28
Dans $\mathbb{R}^3$, on considère les vecteurs $\mathbf{u} = (-2,3,7)$, $\mathbf{v} = (1,-2,-3)$ et $\mathbf{w} = (-1,-1,6)$. Montrer que

$$\text{Vect}\{\mathbf{u},\mathbf{v},\mathbf{w}\} = \text{Vect}\{\mathbf{u},\mathbf{v}\} = \text{Vect}\{\mathbf{v},\mathbf{w}\} = \text{Vect}\{\mathbf{w},\mathbf{u}\}.$$

Correction
Comme $\mathbf{w} = 3\mathbf{u} + 5\mathbf{v}$, on peut tirer de cette relation l'un des vecteurs en fonction des deux autres, ce qui permet de prouver que les espaces engendrés par deux des trois vecteurs sont égaux à $\text{Vect}\{\mathbf{u},\mathbf{v},\mathbf{w}\}$.

Exercice 1.29
Déterminer le rang dans $\mathbb{R}^2$ de la famille $\mathscr{A} = \{\mathbf{u}_1 = (3,1), \mathbf{u}_2 = (-1,5)\}$. Si le rang de la famille est strictement inférieur au nombre de vecteurs de la famille, on déterminera une ou des relations non triviales entre les vecteurs de la famille. Le vecteur $\mathbf{w} = (1,0)$ appartient-il à $\text{Vect}\{\mathbf{u}_1,\mathbf{u}_2\}$? Si oui, l'exprimer comme combinaison linéaire de $\mathbf{u}_1$ et $\mathbf{u}_2$.

Correction
Comme $\det\left(\begin{smallmatrix} 3 & -1 \\ 1 & 5 \end{smallmatrix}\right) \neq 0$, $\text{rg}(\mathscr{A}) = 2$ et l'on a $\text{rg}(\mathscr{A}) = \text{card}(\mathscr{A})$: les vecteurs $\mathbf{u}_1$ et $\mathbf{u}_2$ sont linéaire indépendants. Pour obtenir l'expression de $\mathbf{w}$ en fonction de $\mathbf{u}_1$ et $\mathbf{u}_2$ on cherche les réels a et b tels que $a\mathbf{u}_1 + b\mathbf{u}_2 = \mathbf{w}$, ce qui conduit au système linéaire

$$\begin{cases} 3a-b=1, \\ a+5b=0 \end{cases} \iff a = \frac{5}{16},\ b = \frac{-1}{16},$$

d'où la relation $\mathbf{w} = \frac{1}{16}(5\mathbf{u}_1 - \mathbf{u}_2)$.

Exercice 1.30
Déterminer le rang dans $\mathbb{R}^3$ de la famille $\mathscr{A} = \{\mathbf{u}_1 = (-1,1,-3), \mathbf{u}_2 = (1,2,5), \mathbf{u}_3 = (1,7,1)\}$. Si le rang de la famille est strictement inférieur au nombre de vecteurs de la famille, on déterminera une ou des relations non triviales entre les vecteurs de la famille. Le vecteur $\mathbf{w} = (1,0,0)$ appartient-il à $\text{Vect}\{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3\}$? Si oui, l'exprimer comme combinaison linéaire de $\mathbf{u}_1$, $\mathbf{u}_2$ et $\mathbf{u}_3$.

Correction
Comme $\det\left(\begin{smallmatrix} -1 & 1 & 1 \\ 1 & 2 & 7 \\ -3 & 5 & 1 \end{smallmatrix}\right) \neq 0$, $\text{rg}(\mathscr{A}) = 3$ et l'on a $\text{rg}(\mathscr{A}) = \text{card}(\mathscr{A})$: les vecteurs $\mathbf{u}_1$, $\mathbf{u}_2$ et $\mathbf{u}_3$ sont linéairement indépendants, *i.e.* la famille $\mathscr{A}$ est libre. Comme $\text{card}(\mathscr{A}) = \dim(\mathbb{R}^3)$, alors $\text{Vect}(\mathscr{A}) = \mathbb{R}^3$ et $\mathbf{w} \in \text{Vect}(\mathscr{A})$. Pour obtenir l'expression de $\mathbf{w}$ en fonction de $\mathbf{u}_1$, $\mathbf{u}_2$ et $\mathbf{u}_3$ on cherche les réels a, b et c tels que $a\mathbf{u}_1 + b\mathbf{u}_2 + c\mathbf{u}_3 = \mathbf{w}$, ce qui conduit au système linéaire

$$\begin{cases} -a+b+c=1, \\ a+2b+7c=0, \\ -3a+5b+c=0, \end{cases} \iff \begin{cases} -a+b+c=1, \\ 3b+8c=0, \\ 2b-2c=-3, \end{cases} \iff \begin{cases} -a+b+c=1, \\ 3b+8c=0, \\ -\frac{22}{3}c=-\frac{11}{3}, \end{cases} \iff a=-\frac{3}{2},\ b=-1,\ c=\frac{1}{2},$$

d'où la relation $\mathbf{w} = \frac{1}{2}(-3\mathbf{u}_1 - 2\mathbf{u}_2 + \mathbf{u}_3)$.

Exercice 1.31
Déterminer le rang dans $\mathbb{R}^3$ de la famille $\mathscr{A} = \{\mathbf{u}_1 = (1,4,-3), \mathbf{u}_2 = (2,5,3), \mathbf{u}_3 = (-3,0,-3)\}$. Si le rang de la famille est strictement inférieur au nombre de vecteurs de la famille, on déterminera une ou des relations non triviales entre les vecteurs de la famille. Le vecteur $\mathbf{w} = (1,0,0)$ appartient-il à $\mathrm{Vect}(\mathscr{A})$? Si oui, l'exprimer comme combinaison linéaire de $\mathbf{u}_1$, $\mathbf{u}_2$ et $\mathbf{u}_3$.

Correction
Comme $\det\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 0 \\ -3 & 3 & -3 \end{pmatrix} \neq 0$, $\mathrm{rg}(\mathscr{A}) = 3$ et l'on a $\mathrm{rg}(\mathscr{A}) = \mathrm{card}(\mathscr{A})$: les vecteurs $\mathbf{u}_1$, $\mathbf{u}_2$ et $\mathbf{u}_3$ sont linéaire indépendants, *i.e.* $\mathscr{A}$ est une famille libre. Comme $\mathrm{card}(\mathscr{A}) = \dim(\mathbb{R}^3)$ alors $\mathrm{Vect}(\mathscr{A}) = \mathbb{R}^3$ et $\mathbf{w} \in \mathrm{Vect}(\mathscr{A})$. Pour obtenir l'expression de $\mathbf{w}$ en fonction de $\mathbf{u}_1$, $\mathbf{u}_2$ et $\mathbf{u}_3$ on cherche les réels a, b et c tels que $a\mathbf{u}_1 + b\mathbf{u}_2 + c\mathbf{u}_3 = \mathbf{w}$, ce qui conduit au système linéaire

$$\begin{cases} a+2b+3c=1, \\ 4a+5b=0, \\ -3a+3b-3c=0, \end{cases} \iff \begin{cases} a+2b+3c=1, \\ -3b-12c=-4, \\ 9b+6c=3, \end{cases} \iff \begin{cases} a+2b+3c=1, \\ -3b-12c=-4, \\ -30c=9, \end{cases} \iff a=\frac{-1}{6},\ b=\frac{2}{15},\ c=\frac{3}{10},$$

d'où la relation $\mathbf{w} = -\frac{1}{6}\mathbf{u}_1 + \frac{2}{15}\mathbf{u}_2 + \frac{3}{10}\mathbf{u}_3$.

Exercice 1.32
Déterminer le rang dans $\mathbb{R}^3$ de la famille $\mathscr{A} = \{\mathbf{u}_1 = (1,2,3), \mathbf{u}_2 = (3,2,1), \mathbf{u}_3 = (3,3,3), \mathbf{u}_4 = (7,0,-7)\}$. Si le rang de la famille est strictement inférieur au nombre de vecteurs de la famille, on déterminera une ou des relations non triviales entre les vecteurs de la famille.

Correction
Sans faire de calcul on sait que $\mathrm{rg}(\mathscr{A}) \leq 3$ donc au moins un des vecteurs de la famille est combinaison linéaire des autres. Comme $4\mathbf{u}_3 = 3\mathbf{u}_1 + 3\mathbf{u}_2$ donc $\mathrm{Vect}\{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3,\mathbf{u}_4\} = \mathrm{Vect}\{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_4\}$. Comme $2\mathbf{u}_4 = -7\mathbf{u}_1 + 7\mathbf{u}_2$ alors $\mathrm{Vect}\{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_4\} = \mathrm{Vect}\{\mathbf{u}_1,\mathbf{u}_2\}$. Comme $\mathbf{u}_1$ et $\mathbf{u}_2$ ne sont pas colinéaires, ils sont linéairement indépendants et on conclut que $\mathrm{rg}(\mathscr{A}) = 2$.

Exercice 1.33
Déterminer le rang dans $\mathbb{R}^5$ de la famille

$$\mathscr{A} = \{\mathbf{u}_1 = (1,1,1,2,5), \mathbf{u}_2 = (2,1,0,3,4), \mathbf{u}_3 = (-1,0,-1,4,7), \mathbf{u}_4 = (-9,-2,1,-1,9)\}.$$

Si le rang de la famille est strictement inférieur au nombre de vecteurs de la famille, on déterminera une ou des relations non triviales entre les vecteurs de la famille.

Correction
Sans faire de calcul on sait que $1 \leq \mathrm{rg}(\mathscr{A}) \leq 4$. Comme $\mathbf{u}_4 = 3\mathbf{u}_1 - 5\mathbf{u}_2 + 2\mathbf{u}_3$ donc $\mathrm{Vect}\{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3,\mathbf{u}_4\} = \mathrm{Vect}\{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3\}$. Comme $\det\begin{pmatrix} 1 & 2 & -1 \\ 1 & 1 & 0 \\ 1 & 0 & -1 \end{pmatrix} \neq 0$, on conclut que $\mathrm{rg}(\mathscr{A}) = 3$.

Exercice 1.34
Dans $\mathbb{R}^3$, montrer que l'espace vectoriel engendré par les vecteurs

$$\mathbf{u}_1 = (2,3,-1), \quad \mathbf{u}_2 = (1,-1,-2)$$

et l'espace vectoriel engendré par les vecteurs

$$\mathbf{v}_1 = (3,7,0), \quad \mathbf{v}_2 = (5,0,-7)$$

sont les mêmes.

Correction
Pour montrer l'égalité des deux ensembles, on va prouver les deux inclusions réciproques : $\mathrm{Vect}\{\mathbf{v}_1,\mathbf{v}_2\} \subset \mathrm{Vect}\{\mathbf{u}_1,\mathbf{u}_2\}$ et $\mathrm{Vect}\{\mathbf{u}_1,\mathbf{u}_2\} \subset \mathrm{Vect}\{\mathbf{v}_1,\mathbf{v}_2\}$.

① $\mathrm{Vect}\{\mathbf{v}_1,\mathbf{v}_2\} \subset \mathrm{Vect}\{\mathbf{u}_1,\mathbf{u}_2\}$: il suffit de montrer que $\mathbf{v}_1 \in \{\mathbf{u}_1,\mathbf{u}_2\}$ et $\mathbf{v}_2 \in \{\mathbf{u}_1,\mathbf{u}_2\}$, ce qui suit de la remarque $\mathbf{v}_1 = 2\mathbf{u}_1 - \mathbf{u}_2$ et $\mathbf{v}_2 = \mathbf{u}_1 + 3\mathbf{u}_2$.

② $\mathrm{Vect}\{\mathbf{u}_1,\mathbf{u}_2\} \subset \mathrm{Vect}\{\mathbf{v}_1,\mathbf{v}_2\}$: il suffit de montrer que $\mathbf{u}_1 \in \{\mathbf{v}_1,\mathbf{v}_2\}$ et $\mathbf{u}_2 \in \{\mathbf{v}_1,\mathbf{v}_2\}$, ce qui suit de la remarque $7\mathbf{u}_1 = 3\mathbf{v}_1 + \mathbf{v}_2$ et $7\mathbf{u}_2 = -\mathbf{v}_1 + 2\mathbf{v}_2$.

Exercice 1.35
Montrer que l'espace vectoriel engendré par les vecteurs

$$\mathbf{u}_1 = (1,2,-1,3), \quad \mathbf{u}_2 = (2,4,1,-2), \quad \mathbf{u}_3 = (3,6,3,-7)$$

et l'espace vectoriel engendré par les vecteurs

$$\mathbf{v}_1 = (1,2,-4,11), \quad \mathbf{v}_2 = (2,4,-5,14)$$

sont les mêmes.

Correction
On note U l'espace vectoriel engendré par les vecteurs $\mathbf{u}_1$, $\mathbf{u}_2$ et $\mathbf{u}_3$ et V l'espace vectoriel engendré par les vecteurs $\mathbf{v}_1$, $\mathbf{v}_2$.
Remarquons tout d'abord que si $U = V$ alors $\dim(U) = \dim(V) = 2$ donc la famille $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ n'est pas libre.
Pour démontrer que $U = V$ on montre que $U \subset V$ et que $V \subset U$.

⋆ Pour montrer que $U \subset V$ il suffit de montrer que chaque $\mathbf{u}_i$, $i = 1,2,3$, est combinaison linéaire des vecteurs $\mathbf{v}_1$, $\mathbf{v}_2$:

$$\mathbf{u}_1 = a\mathbf{v}_1 + b\mathbf{v}_2 \iff \begin{cases} 1 = a+2b, \\ 2 = 2a+4b, \\ -1 = -4a-5b, \\ 3 = 11a+14b, \end{cases} \iff a = -1,\ b = 1$$

$$\mathbf{u}_2 = a\mathbf{v}_1 + b\mathbf{v}_2 \iff \begin{cases} 2 = a+2b, \\ 4 = 2a+4b, \\ 1 = -4a-5b, \\ -2 = 11a+14b, \end{cases} \iff a = -4,\ b = 3$$

$$\mathbf{u}_3 = a\mathbf{v}_1 + b\mathbf{v}_2 \iff \begin{cases} 3 = a+2b, \\ 6 = 2a+4b, \\ 3 = -4a-5b, \\ -7 = 11a+14b, \end{cases} \iff a = -7,\ b = 5$$

⋆ Pour montrer que $V \subset U$ il suffit de montrer que chaque $\mathbf{v}_i$, $i = 1,2$, est combinaison linéaire des vecteurs $\mathbf{u}_1$, $\mathbf{u}_2$ et $\mathbf{u}_3$:

$$\mathbf{v}_1 = a\mathbf{u}_1 + b\mathbf{u}_2 + c\mathbf{u}_3 \iff \begin{cases} 1 = a+2b+3c, \\ 2 = 2a+4b+6c, \\ -4 = -a+b+3c, \\ 11 = 3a-2b-7c, \end{cases} \iff a = 3+\kappa,\ b = -1-2\kappa,\ c = \kappa,$$

$$\mathbf{v}_2 = a\mathbf{u}_1 + b\mathbf{u}_2 + c\mathbf{u}_3 \iff \begin{cases} 2 = a+2b+3c, \\ 4 = 2a+4b+6c, \\ -5 = -a+b+3c, \\ 14 = 3a-2b-7c, \end{cases} \iff a = 4-\kappa,\ b = -1-2\kappa,\ c = \kappa,$$

Exercice 1.36
Soient $p_0(x) = x+1$, $p_1(x) = x^2+x$ et $p_2(x) = 2x^2+1$ trois polynômes de $\mathbb{R}_2[x]$. Démontrer que $\text{Vect}\{p_0, p_1, p_2\} = \mathbb{R}_2[x]$.

Correction
Méthode 1 : pour prouver l'égalité de deux ensembles A et B, on peut démontrer que $A \subset B$ et que $B \subset A$. Pour démontrer que $A \subset B$, on considère un élément quelconque de A et on démontre qu'il appartient à B.

⋆ Comme $p_0, p_1, p_2 \in \mathbb{R}_2[x]$ qui est un espace vectoriel, toute combinaison linéaire de ces trois polynômes est encore un élément de $\mathbb{R}_2[x]$, par conséquent $\text{Vect}\{p_0, p_1, p_2\} \subset \mathbb{R}_2[x]$.

⋆ $\mathbb{R}_2[x] \subset \text{Vect}\{p_0, p_1, p_2\}$ ssi pour tout $q \in \mathbb{R}_2[x]$ il existe des réels $\lambda_0, \lambda_1, \lambda_2$ tels que $q = \lambda_0 \cdot p_0 + \lambda_1 \cdot p_1 + \lambda_2 \cdot p_2$:

$$q(x) = a + bx + cx^2 \in \text{Vect}\{p_0, p_1, p_2\}$$
$$\iff \exists\, (\lambda_0, \lambda_1, \lambda_2) \in \mathbb{R}^3 \text{ tel que } q = \lambda_0 \cdot p_0 + \lambda_1 \cdot p_1 + \lambda_2 \cdot p_2$$

$$\begin{aligned}
&\iff \exists\,(\lambda_0,\lambda_1,\lambda_2)\in\mathbb{R}^3 \text{ tel que } a+bx+cx^2=\lambda_0(x+1)+\lambda_1(x^2+x)+\lambda_2(2x^2+1)\\
&\iff \exists\,(\lambda_0,\lambda_1,\lambda_2)\in\mathbb{R}^3 \text{ tel que } a+bx+cx^2=(\lambda_0+\lambda_2)+(\lambda_0+\lambda_1)x+(\lambda_1+2\lambda_2)x^2\\
&\iff \exists\,(\lambda_0,\lambda_1,\lambda_2)\in\mathbb{R}^3 \text{ tel que } \begin{cases}\lambda_0+\lambda_2=a,\\ \lambda_0+\lambda_1=b,\\ \lambda_1+2\lambda_2=c.\end{cases}
\end{aligned}$$

Comme $\begin{vmatrix}1&0&1\\1&1&0\\0&1&2\end{vmatrix}=3$, le système est de Cramer et on peut conclure que $\mathbb{R}_2[x]\subset \text{Vect}\{p_0,p_1,p_2\}$. Après résolution du système linéaire on trouve $q=bp_0+(-a+b+c)p_1+(a-b)p_2$.

Méthode 2 : comme $\text{card}(\{p_0,p_1,p_2\})=3=\dim(\mathbb{R}_2[x])$, il suffit de prouver que la famille $\{p_0,p_1,p_2\}$ est libre, *i.e.* "$\lambda_0\cdot p_0+\lambda_1\cdot p_1+\lambda_2\cdot p_2=0 \implies \lambda_0=\lambda_1=\lambda_2=0$" :

$$\begin{aligned}
&\lambda_0\cdot p_0+\lambda_1\cdot p_1+\lambda_2\cdot p_2=0\\
&\iff \lambda_0(x+1)+\lambda_1(x^2+x)+\lambda_2(2x^2+1)=0\\
&\iff (\lambda_0+\lambda_2)+(\lambda_0+\lambda_1)x+(\lambda_1+2\lambda_2)x^2=0\\
&\iff \begin{cases}\lambda_0+\lambda_2=0,\\ \lambda_0+\lambda_1=0,\\ \lambda_1+2\lambda_2=0.\end{cases}
\end{aligned}$$

Comme $\begin{vmatrix}1&0&1\\1&1&0\\0&1&2\end{vmatrix}=3$, le système admet l'unique solution nulle et on peut conclure que $\text{Vect}\{p_0,p_1,p_2\}=\mathbb{R}_2[x]$.

Exercice 1.37
Étudier si la famille

$$\mathscr{F}=\{\mathbf{u}=(2,3),\mathbf{v}=(4,5)\}$$

de l'espace vectoriel $\mathbb{R}^2$ est libre. Si la famille est liée, trouver une relation entre les vecteurs de cette famille.

Correction
On dit qu'une famille $\mathscr{F}=\{\mathbf{u}_1,\dots,\mathbf{u}_p\}$ est libre lorsque

$$\sum_{i=1}^{p} a_i\cdot\mathbf{u}_i=\mathbf{0}_E \qquad\Longrightarrow\qquad a_i=0\ \forall\ i.$$

Ici

$$\sum_{i=1}^{p} a_i\cdot\mathbf{u}_i=\mathbf{0}_E \iff a_1\mathbf{u}+a_2\mathbf{v}=(0,0) \iff \begin{cases}2a_1+4a_2=0,\\ 3a_1+5a_2=0\end{cases} \iff \begin{cases}a_1=0,\\ a_2=0,\end{cases}$$

donc la famille est libre.
(On peut remarquer que, $\mathscr{F}$ étant libre et comme $\text{card}(\mathscr{F})=2$, elle engendre un espace vectoriel de dimension 2. Puisque $\mathscr{F}\subset\mathbb{R}^2$ et $\dim(\mathbb{R}^2)=2$, on conclut que $\text{Vect}(\mathscr{F})=\mathbb{R}^2$.)

Exercice 1.38
Étudier si la famille

$$\mathscr{F}=\{\mathbf{u}=(1,0,1),\mathbf{v}=(2,1,0),\mathbf{w}=(0,-1,2)\}$$

de l'espace vectoriel $\mathbb{R}^3$ est libre. Si la famille est liée, trouver une relation entre les vecteurs de cette famille.

Correction
On dit qu'une famille $\mathscr{F}=\{\mathbf{u}_1,\dots,\mathbf{u}_p\}$ est libre lorsque

$$\sum_{i=1}^{p} a_i\cdot\mathbf{u}_i=\mathbf{0}_E \qquad\Longrightarrow\qquad a_i=0\ \forall\ i.$$

Ici

$$\sum_{i=1}^{p} a_i \cdot \mathbf{u}_i = \mathbf{0}_E \iff a_1\mathbf{u} + a_2\mathbf{v} + a_3\mathbf{w} = (0,0,0) \iff \begin{cases} a_1 + 2a_2 = 0, \\ a_2 - a_3 = 0, \\ a_1 + 2a_3 = 0 \end{cases} \iff \begin{cases} a_1 = -2\kappa, \\ a_2 = \kappa, \\ a_3 = \kappa \end{cases}$$

donc la famille est liée. De plus, en prenant par exemple $\kappa = 1$, on a $\mathbf{w} = 2 \cdot \mathbf{u} - \mathbf{v}$.

Exercice 1.39
Étudier si la famille

$$\mathscr{F} = \left\{ \mathbb{A} = \begin{pmatrix} 3 & 2 & 1 \\ 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}, \mathbb{B} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \mathbb{C} = \mathbb{I}_3 \right\}$$

de l'espace vectoriel $\mathscr{M}_3(\mathbb{R})$ est libre. Si la famille est liée, trouver une relation entre les vecteurs de cette famille.

Correction
On dit qu'une famille $\mathscr{F} = \{\mathbf{u}_1, \ldots, \mathbf{u}_p\}$ est libre lorsque

$$\sum_{i=1}^{p} a_i \cdot \mathbf{u}_i = \mathbf{0}_E \qquad \Longrightarrow \qquad a_i = 0 \ \forall \ i.$$

Ici

$$\sum_{i=1}^{p} a_i \cdot \mathbf{u}_i = \mathbf{0}_E \iff a_1\mathbb{A} + a_2\mathbb{B} + a_3\mathbb{C} = \mathbb{O}_3 \iff \begin{cases} 3a_1 + a_2 + a_3 = 0, \\ 2a_1 + a_2 = 0, \\ a_1 + a_2 = 0, \\ a_1 + a_2 = 0, \\ 2a_1 + a_2 + a_3 = 0, \\ 3a_1 + a_2 = 0, \\ 2a_1 + a_2 = 0, \\ a_1 + a_2 = 0, \\ 3a_1 + a_2 + a_3 = 0. \end{cases} \iff \begin{cases} a_1 = 0, \\ a_2 = 0, \\ a_3 = 0, \end{cases}$$

donc la famille est libre.

Exercice 1.40
Considérons les matrices d'ordre 2 à coefficients réels

$$\mathbb{A} = \begin{pmatrix} 0 & 0 \\ \kappa & 0 \end{pmatrix}, \qquad \mathbb{B} = \begin{pmatrix} 1 & \kappa \\ -2 & 0 \end{pmatrix}, \qquad \mathbb{C} = \begin{pmatrix} \kappa & 1 \\ -1 & 1 \end{pmatrix}.$$

Pour quelles valeurs de $\kappa \in \mathbb{R}$ les trois matrices forment une famille libre ?

Correction
Nous pouvons tout de suite dire que si $\kappa = 0$ alors la famille n'est pas libre.
La famille $\mathscr{F} = \{\mathbb{A}, \mathbb{B}, \mathbb{C}\}$ est libre lorsque

$$\alpha\mathbb{A} + \beta\mathbb{B} + \gamma\mathbb{C} = \mathbb{O}_2 \qquad \Longrightarrow \qquad \alpha = \beta = \gamma = 0.$$

On a

$$\alpha\mathbb{A} + \beta\mathbb{B} + \gamma\mathbb{C} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \Longrightarrow \begin{cases} \beta + \kappa\gamma = 0, \\ \kappa\beta + \gamma = 0, \\ \kappa\alpha - 2\beta - \gamma = 0, \\ \gamma = 0, \end{cases} \Longrightarrow \begin{cases} \beta = 0, \\ \kappa\alpha = 0, \\ \gamma = 0. \end{cases}$$

La famille est libre ssi $\kappa \neq 0$.

Exercice 1.41
Étudier si la famille

$$\mathscr{F} = \left\{ p_0(x) = x^3 + x^2, p_1(x) = x^2 + x, p_2(x) = x + 1, p_3(x) = x^3 + 1 \right\}$$

de l'espace vectoriel $\mathbb{R}_3[x]$ est libre. Si la famille est liée, trouver une relation entre les vecteurs de cette famille.

Correction
On dit qu'une famille $\mathscr{F} = \{\mathbf{u}_1, \dots, \mathbf{u}_p\}$ est libre lorsque

$$\sum_{i=1}^{p} a_i \cdot \mathbf{u}_i = \mathbf{0}_E \qquad \Longrightarrow \qquad a_i = 0 \ \forall\ i.$$

Ici

$$\sum_{i=0}^{n} a_i \cdot \mathbf{u}_i = \mathbf{0}_E \iff a_0 p_0 + a_1 p_1 + a_2 p_2 + a_3 p_3 = 0 \iff (a_2 + a_3) + (a_1 + a_2)x + (a_0 + a_1)x^2 + (a_0 + a_3)x^3 = 0$$

$$\iff \begin{cases} a_2 + a_3 = 0, \\ a_1 + a_2 = 0, \\ a_0 + a_1 = 0, \\ a_0 + a_3 = 0, \end{cases} \iff \begin{cases} a_0 = \kappa, \\ a_1 = -\kappa, \\ a_2 = \kappa, \\ a_3 = -\kappa, \end{cases} \quad \text{pour tout } \kappa \in \mathbb{R}$$

donc la famille est liée. De plus, en prenant par exemple $\kappa = 1$ on a $p_3 = p_0 - p_1 + p_2$.

Exercice 1.42
On considère dans $\mathbb{R}^n$, $n \geq 4$, une famille de 4 vecteurs linéairement indépendants : $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4\}$. Les familles suivantes sont libres ?

① $\{\mathbf{e}_1, 2\mathbf{e}_2, \mathbf{e}_3\}$
② $\{\mathbf{e}_1, \mathbf{e}_3\}$
③ $\{\mathbf{e}_1, 2\mathbf{e}_1 + \mathbf{e}_4, \mathbf{e}_4\}$
④ $\{3\mathbf{e}_1 + \mathbf{e}_3, \mathbf{e}_3, \mathbf{e}_2 + \mathbf{e}_3\}$
⑤ $\{2\mathbf{e}_1 + \mathbf{e}_2, \mathbf{e}_1 - 3\mathbf{e}_2, \mathbf{e}_4, \mathbf{e}_2 - \mathbf{e}_1\}$

Correction
① Oui
② Oui
③ Non
④ Oui car $\alpha(3\mathbf{e}_1 + \mathbf{e}_3) + \beta(\mathbf{e}_3) + \gamma(\mathbf{e}_2 + \mathbf{e}_3) = \mathbf{0} \implies 3\alpha\mathbf{e}_1 + \gamma\mathbf{e}_2 + (\alpha + \beta + \gamma)\mathbf{e}_3 = \mathbf{0}$. Comme $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ sont linéairement indépendants, cela implique $3\alpha = \gamma = \alpha + \beta + \gamma = 0 \implies \alpha = \beta = \gamma = 0$
⑤ Non : $-7(\mathbf{e}_2 - \mathbf{e}_1) = 2(2\mathbf{e}_1 + \mathbf{e}_2) + 3(\mathbf{e}_1 - 3\mathbf{e}_2)$

Exercice 1.43
On considère dans $\mathbb{R}^n$, $n \geq 4$, une famille de 4 vecteurs linéairement indépendants : $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4\}$. Les familles suivantes sont libres ?

① $\{\mathbf{e}_1 + \mathbf{e}_2\}$
② $\{\mathbf{e}_1 + \mathbf{e}_2, \mathbf{e}_2\}$
③ $\{\mathbf{e}_1 + \mathbf{e}_2, \mathbf{e}_1 - \mathbf{e}_2\}$
④ $\{\mathbf{e}_1 + \mathbf{e}_2, \mathbf{e}_1 - \mathbf{e}_2, \mathbf{e}_1 + \mathbf{e}_3, \mathbf{e}_1 - \mathbf{e}_3\}$
⑤ $\{2\mathbf{e}_1 + \mathbf{e}_2, \mathbf{e}_1 - 3\mathbf{e}_2, \mathbf{e}_4, \mathbf{e}_2 - \mathbf{e}_1\}$

Correction
① Oui
② Oui

③ Oui

④ Non

⑤ Non : $-7(\mathbf{e}_2-\mathbf{e}_1)=2(2\mathbf{e}_1+\mathbf{e}_2)+3(\mathbf{e}_1-3\mathbf{e}_2)$

Exercice 1.44

On considère dans $\mathbb{R}^n$ une famille de 4 vecteurs linéairement indépendants : $\{\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3,\mathbf{e}_4\}$. Les familles suivantes sont libres ?

① $\{2\mathbf{e}_1,\mathbf{e}_2,-\mathbf{e}_3\}$

② $\{\mathbf{e}_1,-\mathbf{e}_3\}$

③ $\{\mathbf{e}_1,3\mathbf{e}_1+\mathbf{e}_4,\mathbf{e}_4\}$

④ $\{3\mathbf{e}_1+\mathbf{e}_3,\mathbf{e}_3,\mathbf{e}_2+\mathbf{e}_3\}$

⑤ $\{2\mathbf{e}_1+\mathbf{e}_2,\mathbf{e}_1-3\mathbf{e}_2,\mathbf{e}_4,\mathbf{e}_2-\mathbf{e}_1\}$

Correction

① Oui

② Oui

③ Non

④ Oui

⑤ Non : $-7(\mathbf{e}_2-\mathbf{e}_1)=2(2\mathbf{e}_1+\mathbf{e}_2)+3(\mathbf{e}_1-3\mathbf{e}_2)$

Exercice 1.45

Écrire la base canonique de $\mathbb{R}_3$, la base canonique de $\mathscr{M}_2(\mathbb{R})$ et la base canonique de $\mathbb{R}_2[t]$.

Correction

1. Base canonique de $\mathbb{R}_3$: $\left\{\begin{pmatrix}1\\0\\0\end{pmatrix},\begin{pmatrix}0\\1\\0\end{pmatrix},\begin{pmatrix}0\\0\\1\end{pmatrix}\right\}$
2. Base canonique de $\mathscr{M}_2(\mathbb{R})$: $\left\{\begin{pmatrix}1&0\\0&0\end{pmatrix},\begin{pmatrix}0&1\\0&0\end{pmatrix},\begin{pmatrix}0&0\\1&0\end{pmatrix},\begin{pmatrix}0&0\\0&1\end{pmatrix}\right\}$
3. Base canonique de $\mathbb{R}_2[t]$: $\left\{1,t,t^2\right\}$

Exercice 1.46

Soit V un espace vectoriel et $\mathscr{F}$ une famille libre d'éléments de V. Donner la définition de $\text{card}(\mathscr{F})$ et de $\dim(V)$. Si $\dim(V)=\text{card}(\mathscr{F})$, que peut-on conclure ?

Correction

1. Le cardinal d'une famille est le nombre d'éléments qui la constitue.
2. La dimension d'un espace vectoriel est le nombre d'éléments d'une de ses bases, *i.e.* le cardinal d'une de ses bases.
3. Si la dimension de l'espace vectoriel V coïncide avec le cardinal de la famille $\mathscr{F}$ contenue dans V, alors la famille $\mathscr{F}$ constitue une base de V. Attention : si la famille $\mathscr{F}$ n'est pas libre, on ne peut rien conclure.

Exercice 1.47

Vrai ou Faux ?

① Toute famille génératrice contient une base.

② La dimension d'un espace vectoriel est le nombre de vecteur de cet espace.

③ Toute famille contenant une famille liée est liée.

④ La base de $\mathbb{R}_3[x]$ est $\left\{1,x,x^2,x^3\right\}$.

⑤ Si $E=\text{Vect}\{\mathbf{u},\mathbf{v},\mathbf{w}\}$ et si $\{\mathbf{u},\mathbf{v},\mathbf{w}\}$ est une famille libre, alors $\dim(E)=3$.

⑥ $\text{Vect}\left\{\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_p\right\}=\text{Vect}\left\{\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_{p-1}\right\}$ si et seulement si $\mathbf{u}_p$ est combinaison linéaire de $\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_{p-1}$.

⑦ Soient $\mathbf{u}$, $\mathbf{v}$ et $\mathbf{w}$ trois vecteurs d'un espace vectoriel E. On suppose que deux vecteurs parmi ces trois ne sont pas colinéaires. Alors la famille $\{\mathbf{u},\mathbf{v},\mathbf{w}\}$ est libre.

Correction

① Vrai (dans le sens que d'une famille génératrice on peut extraire une famille libre qui génère le même espace vectoriel).

② Faux. Un espace vectoriel de dimension finie a une infinité de vecteurs. La dimension d'un espace vectoriel est le nombre de vecteur d'une de ces bases.

③ Vrai. La relation non trivial qui lie des vecteurs de la plus petite des deux familles est vraie dans la plus grande.

④ Incorrect. On ne peut pas parler de «la» base de $\mathbb{R}_3[x]$ car il y en a une infinité.

⑤ Vrai. La famille $\{\mathbf{u},\mathbf{v},\mathbf{w}\}$ est une base de E car libre et génératrice de E.

⑥ Vrai.

⑦ Faux. Par exemple si $\mathbf{w}=\mathbf{u}+\mathbf{v}$, deux vecteurs parmi ces trois ne sont pas colinéaires mais la famille $\{\mathbf{u},\mathbf{v},\mathbf{w}\}$ est liée.

Exercice 1.48

Considérons l'ensemble

$$F=\left\{(x,x,y,y)\;\middle|\;x,y\in\mathbb{R}\right\}.$$

1. Montrer que F est un sous-espace vectoriel de $\mathbb{R}^4$.
2. Donner une base de F et sa dimension.

Correction

1. F est un sous-espace vectoriel de $\mathbb{R}^4$ car

$$F=\left\{\begin{pmatrix}\kappa_1\\ \kappa_1\\ \kappa_2\\ \kappa_2\end{pmatrix}\;\middle|\;\kappa_1,\kappa_2\in\mathbb{R}\right\}=\left\{\kappa_1\begin{pmatrix}1\\1\\0\\0\end{pmatrix}+\kappa_2\begin{pmatrix}0\\0\\1\\1\end{pmatrix}\;\middle|\;\kappa_1,\kappa_2\in\mathbb{R}\right\}=\operatorname{Vect}\left\{\begin{pmatrix}1\\1\\0\\0\end{pmatrix},\begin{pmatrix}0\\0\\1\\1\end{pmatrix}\right\}.$$

2. Les deux vecteurs $\mathbf{u}_1=(1,1,0,0)$ et $\mathbf{u}_2=(0,0,1,1)$ constituent une famille génératrice de F. On vérifie aisément que cette famille est libre donc elle est une base de F. Comme $\operatorname{card}(\{\mathbf{u}_1,\mathbf{u}_2\})=2$, alors $\dim(F)=2$.

Exercice 1.49

Considérons l'ensemble

$$F=\left\{a+ax^2+bx^4\;\middle|\;a,b\in\mathbb{R}\right\}.$$

1. Montrer que F est un sous-espace vectoriel de $\mathbb{R}_4[x]$.
2. Donner une base de F et sa dimension.

Correction

1. F est un sous-espace vectoriel de $\mathbb{R}_4[x]$ car

$$F=\left\{a+ax^2+bx^4\;\middle|\;(a,b)\in\mathbb{R}^2\right\}=\left\{a(1+x^2)+bx^4\;\middle|\;(a,b)\in\mathbb{R}^2\right\}=\operatorname{Vect}\left\{1+x^2,x^4\right\}.$$

2. Les deux polynômes $p(x)=1+x^2$ et $q(x)=x^4$ constituent une famille génératrice de F. On vérifie aisément que cette famille est libre donc elle est une base de F. Comme $\operatorname{card}(\{p,q\})=2$, alors $\dim(F)=2$.

Exercice 1.50

Considérons l'ensemble

$$F=\left\{(x,y,z)\in\mathbb{R}^3\;\middle|\;2x+y+z=0\right\}.$$

1. Montrer que F est un sous-espace vectoriel de $\mathbb{R}^3$.
2. Donner une base de F et sa dimension.

Correction

1. F est un sous-espace vectoriel de $\mathbb{R}^3$ car

$$F=\left\{\begin{pmatrix}\kappa_1\\ \kappa_2\\ -2\kappa_1-\kappa_2\end{pmatrix}\;\middle|\;\kappa_1,\kappa_2\in\mathbb{R}\right\}=\left\{\kappa_1\begin{pmatrix}1\\0\\-2\end{pmatrix}+\kappa_2\begin{pmatrix}0\\1\\-1\end{pmatrix}\;\middle|\;\kappa_1,\kappa_2\in\mathbb{R}\right\}=\operatorname{Vect}\left\{\begin{pmatrix}1\\0\\-2\end{pmatrix},\begin{pmatrix}0\\1\\-1\end{pmatrix}\right\}.$$

2. Les deux vecteurs $\mathbf{u}_1 = (1,0,-2)$ et $\mathbf{u}_2 = (0,1,-1)$ constituent une famille génératrice de F. On vérifie aisément que cette famille est libre donc elle est une base de F. Comme $\text{card}(\{\mathbf{u}_1, \mathbf{u}_2\}) = 2$, alors $\dim(F) = 2$.

Exercice 1.51
Considérons l'ensemble

$$F = \left\{ \begin{pmatrix} a & b & 0 \\ b & a & 0 \\ c & 0 & a+b \end{pmatrix} \middle| \, a,b,c \in \mathbb{R} \right\}.$$

1. Montrer que F est un sous-espace vectoriel de $\mathscr{M}_3(\mathbb{R})$.
2. Donner une base de F et sa dimension.

Correction

1. F est un sous-espace vectoriel de $\mathscr{M}_3(\mathbb{R})$ car

$$F = \left\{ \begin{pmatrix} a & b & 0 \\ b & a & 0 \\ c & 0 & a+b \end{pmatrix} \middle| \, (a,b,c) \in \mathbb{R}^3 \right\} = \left\{ \kappa_1 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} + \kappa_2 \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} + \kappa_3 \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \middle| \, \kappa_1, \kappa_2, \kappa_3 \in \mathbb{R} \right\}$$
$$= \text{Vect} \left\{ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \right\}.$$

2. Les trois matrice $\mathbb{I}_3$, $\mathbb{A} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ et $\mathbb{B} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$ constituent une famille génératrice de F. On vérifie s'il s'agit d'une famille libre : on dit qu'une famille $\mathscr{F} = \{\mathbf{u}_1, \ldots, \mathbf{u}_p\}$ est libre lorsque

$$\sum_{i=1}^{p} a_i \cdot \mathbf{u}_i = \mathbf{0}_E \qquad \Longrightarrow \qquad a_i = 0 \; \forall \; i.$$

Ici

$$\sum_{i=1}^{p} a_i \cdot \mathbf{u}_i = \mathbf{0}_E \iff a_1\mathbb{I}_3 + a_2\mathbb{A} + a_3\mathbb{B} = \mathbb{O}_3 \iff \begin{cases} a_1 = 0, \\ a_2 = 0, \\ 0 = 0, \\ a_2 = 0, \\ a_1 = 0, \\ 0 = 0, \\ a_3 = 0, \\ 0 = 0, \\ a_1 + a_2 = 0, \end{cases} \iff \begin{cases} a_1 = 0, \\ a_2 = 0, \\ a_3 = 0, \end{cases}$$

donc la famille $\mathscr{F} = \{\mathbb{I}_3, \mathbb{A}, \mathbb{B}\}$ est libre et est une base de F. Comme $\text{card}(\mathscr{F}) = 3$, alors $\dim(F) = 3$.

Exercice 1.52
Considérons l'ensemble

$$F = \left\{ p \in \mathbb{R}_3[x] \;\middle|\; p(0) = p(1) = 0 \right\}.$$

1. Montrer que F est un sous-espace vectoriel de $\mathbb{R}_3[x]$.
2. Donner une base de F et sa dimension.

Correction

1. F est un sous-espace vectoriel de $\mathbb{R}_3[x]$ car

$$F = \{ x(x-1)(ax+b) \mid a,b \in \mathbb{R} \} = \left\{ a\Big(x^2(x-1)\Big) + b\Big(x(x-1)\Big) \;\middle|\; a,b \in \mathbb{R} \right\} = \text{Vect}\left\{ x^2(x-1), x(x-1) \right\}.$$

Si on n'a pas remarqué que 0 et 1 sont racines des polynômes de F, il suffit de remarquer que

$$\begin{aligned} F &= \{ p \in \mathbb{R}_3[x] \mid p(0) = p(1) = 0 \} \\ &= \{ a + bx + cx^2 + dx^3 \in \mathbb{R}_3[x] \mid a = 0 \text{ et } a + b + c + d = 0 \} \\ &= \{ bx + cx^2 + (-b-c)x^3 \mid b, c \in \mathbb{R} \} \\ &= \{ b(x - x^3) + c(x^2 - x^3) \mid b, c \in \mathbb{R} \} \\ &= \operatorname{Vect} \{ x - x^3, x^2 - x^3 \}. \end{aligned}$$

Par conséquent F est un sous-espace vectoriel de $\mathbb{R}_3[x]$.

2. Les deux polynômes $p(x) = x - x^3$ et $q(x) = x^2 - x^3$ constituent une famille génératrice de F. On montre que la famille $\mathscr{F} = \{ x - x^3, x^2 - x^3 \}$ est une base de l'espace vectoriel F ; en effet

$$\alpha(x - x^3) + \beta(x^2 - x^3) = 0 \quad \forall x \in \mathbb{R} \qquad \Longleftrightarrow \qquad \alpha x + \beta x^2 + (-\alpha - \beta)x^3 = 0 \quad \forall x \in \mathbb{R} \qquad \Longleftrightarrow \qquad \alpha = \beta = 0.$$

Comme $\operatorname{card}(\mathscr{F}) = 2$, alors $\dim(F) = 2$.

Exercice 1.53
Soit $\mathbb{R}_3[t]$ l'espace vectoriel des polynômes de degré au plus 3. Soit $U = \{ p \in \mathbb{R}_3[t] \mid p(-1) = 0 \}$. Montrer que U est un sous-espace vectoriel de $\mathbb{R}_3[t]$ et en donner une base.

Correction
On montre que $U = \operatorname{Vect} \{ \mathbf{e}_1, \ldots, \mathbf{e}_p \}$ où $\mathbf{e}_1, \ldots, \mathbf{e}_p$ sont des éléments de $\mathbb{R}_3[x]$. En effet

$$\begin{aligned} U &= \{ a + bx + cx^2 + dx^3 \in \mathbb{R}_3[x] \mid a - b + c - d = 0 \} \\ &= \{ a + bx + cx^2 + (a - b + c)x^3 \mid a, b, c \in \mathbb{R} \} \\ &= \{ a(1 + x^3) + b(x - x^3) + c(x^2 + x^3) \mid a, b, c \in \mathbb{R} \} \\ &= \operatorname{Vect} \{ 1 + x^3, x - x^3, x^2 + x^3 \}. \end{aligned}$$

Par conséquent U est un sous-espace vectoriel de $\mathbb{R}_3[x]$.
(On peut également en déduire que $\{ 1 + x^3, x - x^3, x^2 + x^3 \}$ est une famille génératrice de U)

Exercice 1.54
Démontrer que l'ensemble

$$F = \{ p \in \mathbb{R}_3[x] \mid p(0) = p'(1) = 0 \}$$

est un sous-espace vectoriel de $\mathbb{R}_3[x]$ et en donner une base.

Correction
On montre que $F = \operatorname{Vect} \{ \mathbf{e}_1, \ldots, \mathbf{e}_p \}$ où $\mathbf{e}_1, \ldots, \mathbf{e}_p$ sont des éléments de $\mathbb{R}_3[x]$. En effet

$$\begin{aligned} F &= \{ p \in \mathbb{R}_3[x] \mid p(0) = p'(1) = 0 \} \\ &= \{ a + bx + cx^2 + dx^3 \in \mathbb{R}_3[x] \mid a = 0 \text{ et } b + 2c + 3d = 0 \} \\ &= \left\{ bx + cx^2 + \frac{-b - 2c}{3} x^3 \;\middle|\; b, c \in \mathbb{R} \right\} \\ &= \{ b(x - x^3/3) + c(x^2 - 2x^3/3) \mid b, c \in \mathbb{R} \} \\ &= \operatorname{Vect} \left\{ x - \frac{1}{3}x^3, x^2 - \frac{2}{3}x^3 \right\}. \end{aligned}$$

Par conséquent F est un sous-espace vectoriel de $\mathbb{R}_3[x]$.
On montre que la famille $\mathscr{F} = \{ x - \frac{1}{3}x^3, x^2 - \frac{2}{3}x^3 \}$ est une base de l'espace vectoriel F ; en effet

$$\alpha(x - x^3/3) + \beta(x^2 - 2x^3/3) = 0 \quad \forall x \in \mathbb{R} \qquad \Longleftrightarrow \qquad \alpha x + \beta x^2 + (-\alpha - 2\beta)x^3/3 = 0 \quad \forall x \in \mathbb{R} \qquad \Longleftrightarrow \qquad \alpha = \beta = 0.$$

Exercice 1.55
Considérons l'espace vectoriel $\mathscr{M}_2(\mathbb{R})$ des matrices carrées d'ordre 2 avec les entrées réelles. Soient a, b et c trois nombres réels quelconques. Considérons le sous-ensemble

$$E = \left\{ \mathbb{A} \in \mathscr{M}_2(\mathbb{R}) \;\middle|\; \mathbb{A} = \begin{pmatrix} a+b & c \\ 2c & -b \end{pmatrix} \right\}.$$

1. Montrer que E est un sous-espace vectoriel de $\mathscr{M}_2(\mathbb{R})$.
2. Donner une base explicite de E.

Correction

1. E est un sous-espace vectoriel de $\mathscr{M}_2(\mathbb{R})$ car

$$\begin{aligned} E &= \left\{ \begin{pmatrix} a+b & c \\ 2c & -b \end{pmatrix} \;\middle|\; (a,b,c) \in \mathbb{R}^3 \right\} = \left\{ a\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + b\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} + c\begin{pmatrix} 0 & 1 \\ 2 & 0 \end{pmatrix} \;\middle|\; a,b,c \in \mathbb{R} \right\} \\ &= \operatorname{Vect}\left\{ \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 2 & 0 \end{pmatrix} \right\}. \end{aligned}$$

2. Les trois matrice $\mathbb{A} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$, $\mathbb{B} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ et $\mathbb{C} = \begin{pmatrix} 0 & 1 \\ 2 & 0 \end{pmatrix}$ constituent une famille génératrice de E. On vérifie s'il s'agit d'une famille libre : on dit qu'une famille $\mathscr{F} = \{\mathbf{u}_1, \dots, \mathbf{u}_p\}$ est libre lorsque

$$\sum_{i=1}^{p} a_i \cdot \mathbf{u}_i = \mathbf{0}_E \qquad \Longrightarrow \qquad a_i = 0 \;\forall\; i.$$

Ici

$$\sum_{i=1}^{p} a_i \cdot \mathbf{u}_i = \mathbf{0}_E \iff a_1\mathbb{A} + a_2\mathbb{B} + a_3\mathbb{C} = \mathbb{O}_2 \iff \begin{cases} a_1 + a_2 = 0, \\ a_3 = 0, \\ 2a_3 = 0, \\ -a_2 = 0, \end{cases} \iff \begin{cases} a_1 = 0, \\ a_2 = 0, \\ a_3 = 0, \end{cases}$$

donc la famille est libre et est une base de E. Comme $\operatorname{card}\{\mathbb{A}, \mathbb{B}, \mathbb{C}\} = 3$, alors $\dim(E) = 3$.

Exercice 1.56
Trouver une base de l'espace engendré par les polynômes dans les deux familles suivantes

1. $W = \{1 + 2x + 3x^2, x + 2x^2, 1 + 2x + 4x^2, 1 + x\}$
2. $W = \{2 + 2x^2, 2 + x - x^2, 3 + x + x^2, 3 + x + 3x^2\}$

Correction

1. Notons

$$w_1(x) = 1 + 2x + 3x^2, \qquad w_2(x) = x + 2x^2, \qquad w_3(x) = 1 + 2x + 4x^2, \qquad w_4(x) = 1 + x.$$

W est une famille non libre si et seulement si

$$\begin{aligned} &\exists (a,b,c,d) \neq (0,0,0,0) \mid a w_1(x) + b w_2(x) + c w_3(x) + d w_4(x) = 0 \iff \\ &\exists (a,b,c,d) \neq (0,0,0,0) \mid (a + c + d) + (2a + b + 2c + d)x + (3a + 2b + 4c)x^2 = 0 \iff \\ &\begin{cases} a \quad +c+d=0, \\ 2a \;+b+2c+d=0, \\ 3a+2b+4c \quad =0 \end{cases} \iff \begin{cases} a \quad +c \;+d=0, \\ \quad b \quad\; -d=0, \\ 2b+c-3d=0 \end{cases} \iff \begin{cases} a \;+c+d=0, \\ \quad b \quad -d=0, \\ \quad\; c-d=0 \end{cases} \iff \\ &(a,b,c,d) = (-2\kappa, \kappa, \kappa, \kappa),\ \kappa \in \mathbb{R}. \end{aligned}$$

Autrement dit $w_4 = 2w_1 - w_2 - w_3$. On a alors

$$\operatorname{Vect}\{w_1, w_2, w_3, w_4\} = \operatorname{Vect}\{w_1, w_2, w_3\}.$$

Vérifions si la famille $\{w_1, w_2, w_3\}$ est libre : on dit qu'une famille $\mathscr{F} = \{\mathbf{u}_1, \dots, \mathbf{u}_p\}$ est libre lorsque

$$\sum_{i=1}^{p} a_i \cdot \mathbf{u}_i = \mathbf{0}_E \qquad \Longrightarrow \qquad a_i = 0 \ \forall\ i.$$

Ici

$$\begin{aligned}\sum_{i=0}^{n} a_i \cdot \mathbf{u}_i = \mathbf{0}_E &\iff a w_1 + b w_2 + c w_3 = 0\\ &\iff (a+c) + (2a+b+2c)x + (3a+2b+4c)x^2 = 0\\ &\iff \begin{cases} a \quad +c=0,\\ 2a \ +b+2c=0,\\ 3a+2b+4c=0 \end{cases} \iff a=b=c=0\end{aligned}$$

donc la famille est libre. Par conséquent $\{w_1, w_2, w_3\}$ est une base de l'espace $\mathrm{Vect}(W)$.

2. Notons

$$w_1(x) = 2+2x^2, \qquad w_2(x) = 2+x-x^2, \qquad w_3(x) = 3+x+x^2, \qquad w_4(x) = 3+x+3x^2.$$

W est une famille non libre si et seulement si

$$\begin{aligned}&\exists (a,b,c,d) \neq (0,0,0,0) \mid a w_1(x) + b w_2(x) + c w_3(x) + d w_4(x) = 0 \iff\\ &\exists (a,b,c,d) \neq (0,0,0,0) \mid (2a+2b+3c+3d) + (b+c+d)x + (2a-b+c+3d)x^2 = 0 \iff\\ &\begin{cases} 2a+2b+3c+3d=0,\\ \quad b \ +c \ +d=0,\\ 2a \ -b \ +c+3d=0, \end{cases} \iff \begin{cases} 2a+2b+3c+3d=0,\\ \quad b \ +c \ +d=0,\\ \quad b+4c+6d=0, \end{cases} \iff \begin{cases} 2a+2b+3c+3d=0,\\ \quad b \ +c \ +d=0,\\ \quad 3c+5d=0, \end{cases} \iff\\ &(a,b,c,d) = (\kappa, 2\kappa, -5\kappa, 3\kappa),\ \kappa \in \mathbb{R}.\end{aligned}$$

Autrement dit $3w_4 = -w_1 - 2w_2 + 5w_3$. On a alors

$$\mathrm{Vect}\{w_1, w_2, w_3, w_4\} = \mathrm{Vect}\{w_1, w_2, w_3\}.$$

Vérifions si la famille $\{w_1, w_2, w_3\}$ est libre : on dit qu'une famille $\mathscr{F} = \{\mathbf{u}_1, \dots, \mathbf{u}_p\}$ est libre lorsque

$$\sum_{i=1}^{p} a_i \cdot \mathbf{u}_i = \mathbf{0}_E \qquad \Longrightarrow \qquad a_i = 0 \ \forall\ i.$$

Ici

$$\begin{aligned}\sum_{i=0}^{n} a_i \cdot \mathbf{u}_i = \mathbf{0}_E &\iff a w_1 + b w_2 + c w_3 = 0\\ &\iff (2a+2b+3c) + (b+c)x + (2a-b+c)x^2 = 0\\ &\iff \begin{cases} 2a+2b+3c=0,\\ \quad b \ +c=0,\\ \quad b+4c=0, \end{cases} \iff a=b=c=0\end{aligned}$$

donc la famille est libre. Par conséquent $\{w_1, w_2, w_3\}$ est une base de l'espace $\mathrm{Vect}(W)$.

Exercice 1.57

Soit

$$\mathscr{F} = \left\{ \mathbb{B}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \mathbb{B}_2 = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, \mathbb{B}_3 = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}, \mathbb{B}_4 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \right\}$$

une famille d'éléments de l'espace vectoriel $E = \left\{ \mathbb{M} \in \mathscr{M}_2(\mathbb{R}) \,\middle|\, \mathbb{M} = \mathbb{M}^T \right\}$.

1. Montrer que $\mathscr{F}$ est génératrice de E, *i.e.* $E = \mathrm{Vect}(\mathscr{F})$;
2. montrer que $\mathscr{F}$ n'est pas libre ;
3. extraire de $\mathscr{F}$ une base de E.

Correction
On remarque que

$$E = \left\{ \begin{pmatrix} a & b \\ b & c \end{pmatrix} \middle| \, a,b,c \in \mathbb{R} \right\}$$

et $\mathbb{B}_i \in E$ pour tout $i = 1,2,3,4$.

1. La famille $\mathscr{F}$ est génératrice de E si $E = \text{Vect}(\mathscr{F})$, *i.e.* s'il existe λ_1, λ_2, λ_3 et $\lambda_4 \in \mathbb{R}$ tels que

$$\begin{pmatrix} a & b \\ b & c \end{pmatrix} = \lambda_1 \mathbb{B}_1 + \lambda_2 \mathbb{B}_2 + \lambda_3 \mathbb{B}_3 + \lambda_4 \mathbb{B}_4, \qquad \forall a,b,c \in \mathbb{R}$$

   i.e. si et seulement si, pour tout $a,b,c \in \mathbb{R}$, le système linéaire suivant admet (au moins) une solution

$$\begin{cases} \lambda_1 + \lambda_2 = a, \\ \lambda_2 + \lambda_3 = b, \\ \lambda_2 + \lambda_3 = b, \\ \lambda_1 + \lambda_3 + \lambda_4 = c. \end{cases}$$

   La matrice augmentée associée à ce système est

$$[\mathbb{A}|\mathbf{b}] = \left(\begin{array}{cccc|c} 1 & 1 & 0 & 0 & a \\ 0 & 1 & 1 & 0 & b \\ 0 & 1 & 1 & 0 & b \\ 1 & 0 & 1 & 1 & c \end{array} \right).$$

   Comme la seconde et la troisième ligne sont égales alors $\text{rg}(\mathbb{A})$ et $\text{rg}([\mathbb{A}|b])$ sont < 4. Puisque $\begin{vmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{vmatrix} \neq 0$, alors $\text{rg}(\mathbb{A}) = \text{rg}([\mathbb{A}|b]) = 3$: le système admet une solution (non unique car le rang n'est pas maximal).

2. $\mathscr{F}$ n'est pas libre car $2\mathbb{B}_4 = \mathbb{B}_1 - \mathbb{B}_2 + \mathbb{B}_3$

3. On en extrait par exemple la famille $\mathscr{B} = \{\mathbb{B}_1, \mathbb{B}_2, \mathbb{B}_3\}$ qui est une base de E car
   (i) $\mathbb{B}_i \in E$ pour tout $i = 1,2,3$,
   (ii) $\text{card}(\mathscr{B}) = \dim(E)$,
   (iii) $\mathscr{B}$ est libre car la combinaison linéaire $\lambda_1 \mathbb{B}_1 + \lambda_2 \mathbb{B}_2 + \lambda_3 \mathbb{B}_3 = \left(\begin{smallmatrix} 0 & 0 \\ 0 & 0 \end{smallmatrix}\right)$ a comme unique solution $\lambda_1 = \lambda_2 = \lambda_3 = 0$.

Exercice 1.58
Soit

$$\mathscr{F} = \left\{ \mathbb{B}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \mathbb{B}_2 = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, \mathbb{B}_3 = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}, \mathbb{B}_4 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \right\}$$

une famille d'éléments de l'espace vectoriel $E = \left\{ \mathbb{M} \in \mathscr{M}_2(\mathbb{R}) \,\middle|\, \mathbb{M} = \mathbb{M}^T \right\}$.

1. Montrer que $\mathscr{F}$ est génératrice de E, *i.e.* $E = \text{Vect}(\mathscr{F})$;
2. montrer que $\mathscr{F}$ n'est pas libre ;
3. extraire de $\mathscr{F}$ une base de E.

Correction
On remarque que

$$E = \left\{ \begin{pmatrix} a & b \\ b & c \end{pmatrix} \middle| \, a,b,c \in \mathbb{R} \right\}$$

et $\mathbb{B}_i \in E$ pour tout $i = 1,2,3,4$.

1. La famille $\mathscr{F}$ est génératrice de E si $E = \text{Vect}(\mathscr{F})$, *i.e.* s'il existe λ_1, λ_2, λ_3 et $\lambda_4 \in \mathbb{R}$ tels que

$$\begin{pmatrix} a & b \\ b & c \end{pmatrix} = \lambda_1 \mathbb{B}_1 + \lambda_2 \mathbb{B}_2 + \lambda_3 \mathbb{B}_3 + \lambda_4 \mathbb{B}_4, \qquad \forall a,b,c \in \mathbb{R}$$

i.e. si et seulement si, pour tout $a, b, c \in \mathbb{R}$, le système linéaire suivant admet (au moins) une solution

$$\begin{cases} \lambda_1 + \lambda_2 = a, \\ \lambda_2 + \lambda_3 = b, \\ \lambda_2 + \lambda_3 = b, \\ \lambda_3 + \lambda_4 = c. \end{cases}$$

La matrice augmentée associée à ce système est

$$[\mathbb{A}|\mathbf{b}] = \left(\begin{array}{cccc|c} 1 & 1 & 0 & 0 & a \\ 0 & 1 & 1 & 0 & b \\ 0 & 1 & 1 & 0 & b \\ 0 & 0 & 1 & 1 & c \end{array}\right).$$

Comme la seconde et la troisième ligne sont égales alors $\mathrm{rg}(\mathbb{A})$ et $\mathrm{rg}([\mathbb{A}|b])$ sont < 4. Puisque $\begin{vmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{vmatrix} \neq 0$, alors $\mathrm{rg}(\mathbb{A}) = \mathrm{rg}([\mathbb{A}|b]) = 3$: le système admet une solution (non unique car le rang n'est pas maximal).

2. $\mathscr{F}$ n'est pas libre car $\mathbb{B}_4 = \mathbb{B}_1 - \mathbb{B}_2 + \mathbb{B}_3$
3. On en extrait par exemple la famille $\mathscr{B} = \{\mathbb{B}_1, \mathbb{B}_2, \mathbb{B}_3\}$ qui est une base de E car
   (i) $\mathbb{B}_i \in E$ pour tout $i = 1, 2, 3$,
   (ii) $\mathrm{card}(\mathscr{B}) = \dim(E)$,
   (iii) $\mathscr{B}$ est libre car la combinaison linéaire $\lambda_1\mathbb{B}_1 + \lambda_2\mathbb{B}_2 + \lambda_3\mathbb{B}_3 = \left(\begin{smallmatrix} 0 & 0 \\ 0 & 0 \end{smallmatrix}\right)$ a comme unique solution $\lambda_1 = \lambda_2 = \lambda_3 = 0$.

Exercice 1.59
Soit

$$\begin{aligned} q_0(x) &= 1 + x + x^2 + x^3, \\ q_1(x) &= x + x^2 + x^3, \\ q_2(x) &= x^2 + x^3, \\ q_3(x) &= x^3, \end{aligned}$$

quatre polynômes de $\mathbb{R}_3[x]$.

1. Démontrer que l'ensemble $\{q_0, q_1, q_2, q_3\}$ est une base de $\mathbb{R}_3[x]$ qu'on notera $\mathscr{B}$.
2. Notons $\mathscr{C} = \{c_0, c_1, c_2, c_3\}$ la base canonique de $\mathbb{R}_3[x]$. Calculer $\mathrm{coord}(c_i, \mathscr{B})$ et $\mathrm{coord}(q_i, \mathscr{C})$.
3. Exprimer le polynôme $a + bx + cx^2 + dx^3$ dans la base $\mathscr{B}$.

Correction

1. Pour montrer que l'ensemble $\{q_0, q_1, q_2, q_3\}$ est une base de $\mathbb{R}_3[x]$ il faut montrer qu'il s'agit d'une famille libre et génératrice de $\mathbb{R}_3[x]$. On dit qu'une famille $\mathscr{F} = \{\mathbf{u}_1, \ldots, \mathbf{u}_p\}$ est libre lorsque

$$\sum_{i=1}^{p} a_i \cdot \mathbf{u}_i = \mathbf{0}_E \qquad \Longrightarrow \qquad a_i = 0 \ \forall\ i = 1, \ldots, p.$$

Ici

$$\sum_{i=0}^{n} a_i \cdot \mathbf{u}_i = \mathbf{0}_E \iff a_0 q_0 + a_1 q_1 + a_2 q_2 + a_3 q_3 = 0$$

$$\iff a_0 + (a_0 + a_1)x + (a_0 + a_1 + a_2)x^2 + (a_0 + a_1 + a_2 + a_3)x^3 = 0 \iff \begin{cases} a_0 = 0, \\ a_1 = 0, \\ a_2 = 0, \\ a_3 = 0, \end{cases}$$

donc la famille est libre. Comme $\mathrm{card}(\mathrm{Vect}\{q_0, q_1, q_2, q_3\}) = 4$ et $\dim(\mathbb{R}_4[x]) = 4$, alors $\mathscr{B} = \{q_0, q_1, q_2, q_3\}$ est une base de $\mathbb{R}_3[x]$.

2. Soit $\mathscr{C} = \{c_0(x) = 1, c_1(x) = x, c_2(x) = x^2, x_3(x) = x^3\}$ la base canonique de $\mathbb{R}_3[x]$. Calculons les coordonnées de q_j

dans la base $\mathscr{C}$:

$$\begin{aligned} q_0(x) &= 1\cdot c_0(x)+1\cdot c_1(x)+1\cdot c_2(x)+1\cdot c_3(x) &&\Longrightarrow \text{coord}(q_0,\mathscr{C})=(1,1,1,1)\\ q_1(x) &= 0\cdot c_0(x)+1\cdot c_1(x)+1\cdot c_2(x)+1\cdot c_3(x) &&\Longrightarrow \text{coord}(q_1,\mathscr{C})=(0,1,1,1)\\ q_2(x) &= 0\cdot c_0(x)+0\cdot c_1(x)+1\cdot c_2(x)+1\cdot c_3(x) &&\Longrightarrow \text{coord}(q_2,\mathscr{C})=(0,0,1,1)\\ q_3(x) &= 0\cdot c_0(x)+0\cdot c_1(x)+0\cdot c_2(x)+1\cdot c_3(x) &&\Longrightarrow \text{coord}(q_3,\mathscr{C})=(0,0,0,1) \end{aligned}$$

En résolvant le système linéaire (ici c'est très facile car il s'agit d'un système triangulaire), on obtient

$$\begin{cases} q_0 &= c_0+c_1+c_2+c_3,\\ q_1 &= c_1+c_2+c_3,\\ q_2 &= c_2+c_3,\\ q_3 &= c_3, \end{cases} \Longleftrightarrow \begin{cases} c_0 &= q_0-q_1,\\ c_1 &= q_1-q_2,\\ c_2 &= q_2-q_3,\\ c_3 &= q_3, \end{cases}$$

Cela signifie que

$$\begin{aligned} c_0(x) &= 1\cdot q_0(x)-1\cdot q_1(x)+0\cdot q_2(x)+0\cdot q_3(x) && \textit{i.e.}\quad \text{coord}(c_0,\mathscr{B})=(1,-1,0,0),\\ c_1(x) &= 0\cdot q_0(x)+1\cdot q_1(x)-1\cdot q_2(x)+0\cdot q_3(x) && \textit{i.e.}\quad \text{coord}(c_1,\mathscr{B})=(0,1,-1,0),\\ c_2(x) &= 0\cdot q_0(x)+0\cdot q_1(x)+1\cdot q_2(x)-1\cdot q_3(x) && \textit{i.e.}\quad \text{coord}(c_2,\mathscr{B})=(0,0,1,-1),\\ c_3(x) &= 0\cdot q_0(x)+0\cdot q_1(x)+0\cdot q_2(x)+1\cdot q_3(x) && \textit{i.e.}\quad \text{coord}(c_3,\mathscr{B})=(0,0,0,1). \end{aligned}$$

3. Soit le polynôme $p(x)=a+bx+cx^2+dx^3$; dans la base $\mathscr{C}$ il a coordonnées (a,b,c,d), donc

$$p(x)=ac_0+bc_1+cc_2+dc_3=a(q_0-q_1)+b(q_1-q_2)+c(q_2-q_3)+d(q_3)=aq_0+(b-a)q_1+(c-b)q_2+(d-c)q_3.$$

Par conséquent, dans la base $\mathscr{B}$ le polynôme $p(x)=a+bx+cx^2+dx^3$ a coordonnées

$$\text{coord}(p,\mathscr{B})=(a,b-a,c-b,d-c).$$

Exercice 1.60 (Rang d'une famille de vecteurs)
Dans $\mathbb{R}^3$, déterminer le rang de la famille

$$\mathscr{E}=\{\mathbf{u}_1=(1,2,3),\mathbf{u}_2=(2,-1,1),\mathbf{u}_3=(1,0,1),\mathbf{u}_4=(0,1,1)\}.$$

Correction
Notons $\mathscr{F}=\text{Vect}(\mathscr{E})$ le sous-espace vectoriel engendré par la famille $\mathscr{E}$.

- ⋆ Comme $\text{card}(\mathscr{E})=4$ alors $\text{rg}(\mathscr{E})=\dim(\mathscr{F})\leq 4$.
- ⋆ Comme $\mathscr{F}$ est un sous-espace vectoriel de $\mathbb{R}^3$, $\dim(\mathscr{F})\leq\dim(\mathbb{R}^3)=3$ ainsi $\text{rg}(\mathscr{E})=\dim(\mathscr{F})\leq 3$.
- ⋆ Comme $\mathbf{u}_1$ et $\mathbf{u}_2$ sont linéairement indépendants, alors $\dim(\text{Vect}\{\mathbf{u}_1,\mathbf{u}_2\})=2$ et comme $\text{Vect}\{\mathbf{u}_1,\mathbf{u}_2\}\subset\mathscr{F}$, on obtient $\text{rg}(\mathscr{E})=\dim(\mathscr{F})\geq 2$.
- ⋆ Étudions maintenant la famille $\{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3\}\subset\mathscr{E}$: si elle est libre, comme $\dim(\text{Vect}\{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3\})=3$ alors $\text{rg}(\mathscr{E})=3$; si elle est liée on ne peut pas conclure. On étudiera alors la famille $\{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_4\}\subset\mathscr{E}$: si elle est libre, comme $\dim(\text{Vect}\{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_4\})=3$ alors $\text{rg}(\mathscr{E})=3$; si elle est liée $\text{rg}(\mathscr{E})=2$.
  Comme $5\mathbf{u}_3=\mathbf{u}_1+2\mathbf{u}_2$, la famille $\{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3\}$ est liée.
  Comme $5\mathbf{u}_4=2\mathbf{u}_1-\mathbf{u}_2$, la famille $\{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_4\}$ est liée.

On conclut que $\text{rg}(\mathscr{E})=2$.

Exercice 1.61
Soit $\mathbb{R}_3[t]$ l'espace des polynômes de degré au plus 3 et considérons l'ensemble

$$V=\left\{p\in\mathbb{R}_3[t]\;\middle|\;p(0)+p(2)=0,\ p(1)=3p(-1)\right\}.$$

1. Montrer que V est un sous-espace vectoriel de $\mathbb{R}_3[t]$.
2. Déterminer une base et la dimension de V.

3. Montrer que le polynôme $p(t) = 2+2t-t^3$ est dans V et trouver les composantes de p dans la base de V calculée auparavant.

Correction

1. On montre que $V = \text{Vect}\{\mathbf{e}_1,\dots,\mathbf{e}_p\}$ où $\mathbf{e}_1,\dots,\mathbf{e}_p$ sont des éléments de $\mathbb{R}_3[t]$. En effet

$$\begin{aligned} V &= \{ p \in \mathbb{R}_3[t] \mid p(0)+p(2)=0,\ p(1)=3p(-1)\} \\ &= \{ a+bt+ct^2+dt^3 \in \mathbb{R}_3[t] \mid 2a+2b+4c+8d=0 \text{ et } a+b+c+d=3a-3b+3c-3d\} \\ &= \left\{ \left(-\frac{5}{3}c-2d\right)+\left(-\frac{1}{3}c-2d\right)t+ct^2+dt^3 \;\middle|\; c,d\in\mathbb{R}\right\} \\ &= \left\{ \left(-\frac{5}{3}-\frac{1}{3}t+t^2\right)c+\left(-2-2t+t^3\right)d \;\middle|\; c,d\in\mathbb{R}\right\} \\ &= \text{Vect}\left\{ q_1(t)=-\frac{5}{3}-\frac{1}{3}t+t^2; q_2(t)=-2-2t+t^3\right\}. \end{aligned}$$

   Par conséquent V est un sous-espace vectoriel de $\mathbb{R}_3[t]$.

2. La famille $\mathcal{V} = \{q_1, q_2\}$ est génératrice de l'espace vectoriel V. De plus,

$$\alpha q_1+\beta q_2 = 0_{\mathbb{R}_3[t]} \quad\Longleftrightarrow\quad \alpha q_1(t)+\beta q_2(t)=0\ \forall t\in\mathbb{R} \quad\Longleftrightarrow\quad \alpha=\beta=0$$

   donc elle est aussi libre donc elle est une base de V et $\dim(V)=2$.

3. Si $p(t)=2+2t-t^3$ on a $p(0)+p(2)=(2)+(2+4-8)=0$ et $p(1)-3p(-1)=(2+2-1)-3(2-2+1)=0$ donc $p\in V$ et $\text{coord}(p,\mathcal{V})=(0,-1)$.

Exercice 1.62

Montrer que les matrices

$$\mathbb{A}=\begin{pmatrix}1&-2\\-2&1\end{pmatrix} \qquad \mathbb{B}=\begin{pmatrix}1&3\\3&6\end{pmatrix} \qquad \mathbb{C}=\begin{pmatrix}-1&1\\1&-3\end{pmatrix}$$

forment un base de l'espace vectoriel M des matrices symétriques d'ordre 2. Décomposer sur cette base la matrice

$$\mathbb{G}=\begin{pmatrix}5&6\\6&7\end{pmatrix}.$$

Correction

La famille $\mathcal{F} = \{\mathbb{A},\mathbb{B},\mathbb{C}\}$ est une base de l'espace vectoriel

$$M = \{\mathbb{M}\in\mathcal{M}_2(\mathbb{R}) \mid \mathbb{M}=\mathbb{M}^T\} = \left\{\begin{pmatrix}a&b\\b&c\end{pmatrix} \mid a,b,c\in\mathbb{R}\right\}$$

car

(i) $\mathbb{A},\mathbb{B},\mathbb{C}\in M$,

(ii) $\text{card}(\mathcal{F})=\dim(M)$,

(iii) $\mathcal{F}$ est libre car la combinaison linéaire $\lambda_1\mathbb{A}+\lambda_2\mathbb{B}+\lambda_3\mathbb{C}=\begin{pmatrix}0&0\\0&0\end{pmatrix}$ a comme unique solution $\lambda_1=\lambda_2=\lambda_3=0$:

$$\lambda_1\begin{pmatrix}1&-2\\-2&1\end{pmatrix}+\lambda_2\begin{pmatrix}1&3\\3&6\end{pmatrix}+\lambda_3\begin{pmatrix}-1&1\\1&-3\end{pmatrix}=\begin{pmatrix}0&0\\0&0\end{pmatrix} \Longleftrightarrow \begin{pmatrix}\lambda_1+\lambda_2-\lambda_3 & -2\lambda_1+3\lambda_2+\lambda_3\\ -2\lambda_1+3\lambda_2+\lambda_3 & \lambda_1+6\lambda_2-3\lambda_3\end{pmatrix}=\begin{pmatrix}0&0\\0&0\end{pmatrix}$$

$$\Longleftrightarrow \begin{cases}\lambda_1+\lambda_2-\lambda_3=0,\\ -2\lambda_1+3\lambda_2+\lambda_3=0,\\ -2\lambda_1+3\lambda_2+\lambda_3=0,\\ \lambda_1+6\lambda_2-3\lambda_3=0,\end{cases} \Longleftrightarrow \begin{pmatrix}1&1&-1\\-2&3&3\\-2&3&3\\1&6&-3\end{pmatrix}\begin{pmatrix}\lambda_1\\\lambda_2\\\lambda_3\end{pmatrix}=\begin{pmatrix}0\\0\\0\\0\end{pmatrix} \Longleftrightarrow \begin{pmatrix}1&1&-1\\-2&3&3\\1&6&-3\end{pmatrix}\begin{pmatrix}\lambda_1\\\lambda_2\\\lambda_3\end{pmatrix}=\begin{pmatrix}0\\0\\0\end{pmatrix}$$

$$\Longleftrightarrow \begin{pmatrix}1&1&-1\\0&5&1\\0&5&-2\end{pmatrix}\begin{pmatrix}\lambda_1\\\lambda_2\\\lambda_3\end{pmatrix}=\begin{pmatrix}0\\0\\0\end{pmatrix} \Longleftrightarrow \begin{pmatrix}1&1&-1\\0&5&1\\0&0&-3\end{pmatrix}\begin{pmatrix}\lambda_1\\\lambda_2\\\lambda_3\end{pmatrix}=\begin{pmatrix}0\\0\\0\end{pmatrix} \Longleftrightarrow \begin{pmatrix}\lambda_1\\\lambda_2\\\lambda_3\end{pmatrix}=\begin{pmatrix}0\\0\\0\end{pmatrix}.$$

On cherche maintenant $\text{coord}(\mathbb{G},\mathcal{F})$ les coordonnées de la matrice $\mathbb{G}$ dans la base $\mathcal{F}$, *i.e.* les uniques $\lambda_1,\lambda_2,\lambda_3\in\mathbb{R}$ tels que $\lambda_1\mathbb{A}+\lambda_2\mathbb{B}+\lambda_3\mathbb{C}=\mathbb{G}$:

$$\lambda_1\begin{pmatrix}1&-2\\-2&1\end{pmatrix}+\lambda_2\begin{pmatrix}1&3\\3&6\end{pmatrix}+\lambda_3\begin{pmatrix}-1&1\\1&-3\end{pmatrix}=\begin{pmatrix}5&6\\6&7\end{pmatrix}\iff\begin{pmatrix}\lambda_1+\lambda_2-\lambda_3&-2\lambda_1+3\lambda_2+\lambda_3\\-2\lambda_1+3\lambda_2+\lambda_3&\lambda_1+6\lambda_2-3\lambda_3\end{pmatrix}=\begin{pmatrix}5&6\\6&7\end{pmatrix}$$

$$\iff\begin{cases}\lambda_1+\lambda_2-\lambda_3=5,\\-2\lambda_1+3\lambda_2+\lambda_3=6,\\-2\lambda_1+3\lambda_2+\lambda_3=6,\\\lambda_1+6\lambda_2-3\lambda_3=7,\end{cases}\iff\begin{pmatrix}1&1&-1\\-2&3&3\\-2&3&3\\1&6&-3\end{pmatrix}\begin{pmatrix}\lambda_1\\\lambda_2\\\lambda_3\end{pmatrix}=\begin{pmatrix}5\\6\\6\\7\end{pmatrix}\iff\begin{pmatrix}1&1&-1\\-2&3&3\\1&6&-3\end{pmatrix}\begin{pmatrix}\lambda_1\\\lambda_2\\\lambda_3\end{pmatrix}=\begin{pmatrix}5\\6\\7\end{pmatrix}$$

$$\iff\begin{pmatrix}1&1&-1\\0&5&1\\0&5&-2\end{pmatrix}\begin{pmatrix}\lambda_1\\\lambda_2\\\lambda_3\end{pmatrix}=\begin{pmatrix}5\\16\\2\end{pmatrix}\iff\begin{pmatrix}1&1&-1\\0&5&1\\0&0&-3\end{pmatrix}\begin{pmatrix}\lambda_1\\\lambda_2\\\lambda_3\end{pmatrix}=\begin{pmatrix}5\\16\\-14\end{pmatrix}\iff\begin{pmatrix}\lambda_1\\\lambda_2\\\lambda_3\end{pmatrix}=\begin{pmatrix}37/5\\34/15\\14/3\end{pmatrix}=\text{coord}(\mathbb{G},\mathscr{F}).$$

1.5.3. Systèmes linéaires : utilisation de fonctions prédéfinies dans Octave

Exercice 1.63 (Systèmes linéaires)
On a demandé à 215 étudiants de dire quel est leur langage de programmation préféré parmi Python, Matlab, Java et C. Chaque étudiant devait fournir une seule réponse. On sait que 163 étudiants ont déclaré préférer Python ou Matlab ; 65 ont déclaré préférer Matlab ou C ; 158 ont déclaré préférer Python ou C.
Traduire les données par un système linéaire de 4 équations et 4 inconnues et le résoudre par la méthode de Gauss.

Correction
On note p, m, j et c le nombre d'étudiants préférant respectivement Python, Matlab, Java et C. D'après l'énoncé on a

$$\begin{cases}p+m+j+c=215,\\p+m=163,\\m+c=65,\\p+c=158.\end{cases}$$

On passe à la notation matricielle et on résout le système par la méthode de Gauss (4 équations donc 3 étapes) :

$$\left(\begin{array}{cccc|c}1&1&1&1&215\\1&1&0&0&163\\0&1&0&1&65\\1&0&0&1&158\end{array}\right)\xrightarrow[\substack{L_3\leftarrow L_3\\L_4\leftarrow L_4-L_1}]{L_2\leftarrow L_2-L_1}{\scriptstyle\text{Étape 1}}\left(\begin{array}{cccc|c}1&1&1&1&215\\0&0&-1&-1&-52\\0&1&0&1&65\\0&-1&-1&0&-57\end{array}\right)\xrightarrow[\text{Pivot nul}]{L_2\leftrightarrow L_3}\left(\begin{array}{cccc|c}1&1&1&1&215\\0&1&0&1&65\\0&0&-1&-1&-52\\0&-1&-1&0&-57\end{array}\right)$$

$$\xrightarrow[\substack{L_3\leftarrow L_3\\L_4\leftarrow L_4+L_2}]{}{\scriptstyle\text{Étape 2}}\left(\begin{array}{cccc|c}1&1&1&1&215\\0&1&0&1&65\\0&0&-1&-1&-52\\0&0&-1&1&8\end{array}\right)\xrightarrow{L_4\leftarrow L_4-L_3}{\scriptstyle\text{Étape 3}}\left(\begin{array}{cccc|c}1&1&1&1&215\\0&1&0&1&65\\0&0&-1&-1&-52\\0&0&0&2&60\end{array}\right).$$

On résout le système triangulaire supérieur ainsi obtenu par remonté :

$$\begin{aligned}2c=60&\implies c=30,\\-j-c=-52&\implies j=-c+52=22,\\m+c=65&\implies m=-c+65=35,\\p+m+j+c=215&\implies p=215-m-j-c=128.\end{aligned}$$

Les préférences sont donc : 128 pour Python, 35 pour Matlab, 22 pour Java et 30 pour C.

```
A=[1 1 1 1; 1 1 0 0; 0 1 0 1; 1 0 0 1]
b=[215; 163; 65; 158]
sol=A\b
```

Systèmes linéaires : méthode de Gauss pour des systèmes carrés

Exercice 1.64
Résoudre les systèmes linéaires suivants :

$$①\ \begin{cases} x_1+2x_2-x_3=2 \\ x_1-2x_2-3x_3=-6 \\ x_1+4x_2+4x_3=3 \end{cases} \qquad ②\ \begin{cases} -x_1+x_2+3x_3=12 \\ 2x_1-x_2+2x_3=-8 \\ 4x_1+x_2-4x_3=15 \end{cases} \qquad ③\ \begin{cases} -2u-4v+3w=-1 \\ 2v-w=1 \\ u+v-3w=-6 \end{cases}$$

$$④\ \begin{cases} -2x-y+4t=2 \\ 2x+3y+3z+2t=14 \\ x+2y+z+t=7 \\ -x-z+t=-1 \end{cases} \qquad ⑤\ \begin{pmatrix} 6 & 1 & 1 \\ 2 & 4 & 0 \\ 1 & 2 & 6 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}=\begin{pmatrix} 12 \\ 0 \\ 6 \end{pmatrix} \qquad ⑥\ \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \\ 3 & 4 & 1 & 2 \\ 4 & 1 & 2 & 3 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}=\begin{pmatrix} 10 \\ 10 \\ 10 \\ 10 \end{pmatrix}$$

Correction
On utilise la méthode du pivot de GAUSS :

①

$$\begin{cases} x_1+2x_2-x_3=2, \\ x_1-2x_2-3x_3=-6, \\ x_1+4x_2+4x_3=3. \end{cases} \xrightarrow[L_3\leftarrow L_3-L_1]{L_2\leftarrow L_2-L_1} \begin{cases} x_1+2x_2-x_3=2, \\ -4x_2-2x_3=-8, \\ 2x_2+5x_3=1. \end{cases} \xrightarrow{L_3\leftarrow L_3+L_2/2} \begin{cases} x_1+2x_2-x_3=2, \\ -4x_2-2x_3=-8, \\ 4x_3=-3. \end{cases}$$

donc $x_3=\frac{-3}{4}$, $x_2=\frac{19}{8}$ et $x_1=\frac{-7}{2}$.

②

$$\begin{cases} -x_1+x_2+3x_3=12 \\ 2x_1-x_2+2x_3=-8 \\ 4x_1+x_2-4x_3=15 \end{cases} \xrightarrow[L_3\leftarrow L_3+4L_1]{L_2\leftarrow L_2+2L_1} \begin{cases} -x_1+x_2+3x_3=12 \\ x_2+8x_3=16 \\ 5x_2+8x_3=63 \end{cases} \xrightarrow{L_3\leftarrow L_3-5L_2} \begin{cases} -x_1+x_2+3x_3=12 \\ x_2+8x_3=16 \\ -32x_3=-17 \end{cases}$$

donc $x_3=\frac{17}{32}$, $x_2=\frac{47}{4}$ et $x_1=\frac{43}{32}$.

③

$$\begin{cases} -2u-4v+3w=-1 \\ 2v-w=1 \\ u+v-3w=-6 \end{cases} \xrightarrow{L_3\leftarrow L_3+L_1/2} \begin{cases} -2u-4v+3w=-1 \\ 2v-w=1 \\ -v-\frac{3}{2}w=-{}^{13}/_{2} \end{cases} \xrightarrow{L_3\leftarrow L_3+L_2/2} \begin{cases} -2u-4v+3w=-1 \\ 2v-w=1 \\ -2w=-6 \end{cases}$$

donc $w=3$, $v=2$ et $u=1$.

④

$$\begin{cases} -2x-y+4t=2 \\ 2x+3y+3z+2t=14 \\ x+2y+z+t=7 \\ -x-z+t=-1 \end{cases} \xrightarrow[L_4\leftarrow L_4-L_1/2]{L_2\leftarrow L_2+L_1 \\ L_3\leftarrow L_3+L_1/2} \begin{cases} -2x-y+4t=2 \\ 2y+3z+6t=16 \\ \frac{3}{2}y+z+3t=8 \\ \frac{1}{2}y-z-t=-2 \end{cases}$$

$$\xrightarrow[L_4\leftarrow L_2-L_2/4]{L_3\leftarrow L_3-3L_2/4} \begin{cases} -2x-y+4t=2 \\ 2y+3z+6t=16 \\ -\frac{5}{4}z-\frac{3}{2}t=-4 \\ -\frac{7}{4}z-\frac{5}{2}t=-6 \end{cases} \xrightarrow{L_4\leftarrow L_4-7L_3/5} \begin{cases} -2x-y+4t=2 \\ 2y+3z+6t=16 \\ -\frac{5}{4}z-\frac{3}{2}t=-4 \\ -\frac{2}{5}t=-\frac{2}{5} \end{cases}$$

donc $t=1$, $z=2$, $y=2$ et $x=0$.

⑤

$$[\mathbb{A}|\mathbf{b}]=\left(\begin{array}{ccc|c} 6 & 1 & 1 & 12 \\ 2 & 4 & 0 & 0 \\ 1 & 2 & 6 & 6 \end{array}\right) \xrightarrow[L_3\leftarrow L_3-\frac{1}{6}L_1]{L_2\leftarrow L_2-\frac{2}{6}L_1} \left(\begin{array}{ccc|c} 6 & 1 & 1 & 12 \\ 0 & \frac{11}{3} & -\frac{1}{3} & -4 \\ 0 & \frac{11}{6} & \frac{35}{6} & 4 \end{array}\right) \xrightarrow{L_3\leftarrow L_3-\frac{\frac{11}{6}}{\frac{11}{3}}L_2} \left(\begin{array}{ccc|c} 6 & 1 & 1 & 12 \\ 0 & \frac{11}{3} & -\frac{1}{3} & -4 \\ 0 & 0 & 6 & 6 \end{array}\right)$$

donc

$$\begin{cases} 6x_1+x_2+x_3=12, \\ \frac{11}{3}x_2-\frac{1}{3}x_3=-4 \\ 6x_3=6 \end{cases} \quad\Longrightarrow\quad x_3=1,\quad x_2=-1,\quad x_1=2.$$

⑥

$$[\mathbb{A}|\mathbf{b}] = \left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 10 \\ 2 & 3 & 4 & 1 & 10 \\ 3 & 4 & 1 & 2 & 10 \\ 4 & 1 & 2 & 3 & 10 \end{array}\right) \xrightarrow[L_4 \leftarrow L_4 - 4L_1]{\substack{L_2 \leftarrow L_2 - 2L_1 \\ L_3 \leftarrow L_3 - 3L_1}} \left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 10 \\ 0 & -1 & -2 & -7 & -10 \\ 0 & -2 & -8 & -10 & -20 \\ 0 & -7 & -10 & -13 & -30 \end{array}\right)$$

$$\xrightarrow[L_4 \leftarrow L_4 - 7L_2]{\substack{L_3 \leftarrow L_3 - 2L_2 \\ L_4 \leftarrow L_4 - 7L_2}} \left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 10 \\ 0 & -1 & -2 & -7 & -10 \\ 0 & 0 & -4 & 4 & 0 \\ 0 & 0 & 4 & 36 & 40 \end{array}\right) \xrightarrow{L_4 \leftarrow L_4 + L_3} \left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 10 \\ 0 & -1 & -2 & -7 & -10 \\ 0 & 0 & -4 & 4 & 0 \\ 0 & 0 & 0 & 40 & 40 \end{array}\right)$$

donc

$$\begin{cases} x_1 + 2x_2 + 3x_3 + 4x_4 = 10 \\ -x_2 - 2x_3 - 7x_4 = -10 \\ -4x_3 + 4x_4 = 0 \\ 40x_4 = 40 \end{cases} \quad \Longrightarrow \quad x_4 = 1, \quad x_3 = 1, \quad x_2 = 1, \quad x_1 = 1.$$

```
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 1.65
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

$$\begin{cases} 2x_1 - x_2 - 3x_3 = 0, \\ -x_1 + 2x_3 = 0, \\ 2x_1 - 3x_2 - x_3 = 0. \end{cases} \xrightarrow[L_3 \leftarrow L_3 - L_1]{L_2 \leftarrow L_2 + L_1/2} \begin{cases} 2x_1 - x_2 - 3x_3 = 0, \\ -{}^1\!/_2 x_2 + {}^1\!/_2 x_3 = 0, \\ -2x_2 + 2x_3 = 0, \end{cases} \xrightarrow{L_3 \leftarrow L_3 - 4L_2} \begin{cases} 2x_1 - x_2 - 3x_3 = 0, \\ -{}^1\!/_2 x_2 + {}^1\!/_2 x_3 = 0, \\ 0 = 0, \end{cases}$$

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 $\mathrm{rg}(S) < 3$. Il admet une infinité de solutions de la forme $(2\kappa, \kappa, \kappa)$, $\kappa \in \mathbb{R}$.
Que dit Octave ?

```
>> [2,-1,-3;-1,0,2;2,-3,-1]\[0;0;0]
warning: matrix singular to machine precision
```

Exercice 1.66
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 :

$$\mathbb{A}=\begin{pmatrix}-3&1&2\\-2&0&2\\-11&6&5\end{pmatrix}\xrightarrow[L_3\leftarrow L_3-11L_1/3]{L_2\leftarrow L_2-2L_1/3}\begin{pmatrix}-3&1&2\\0&-2/3&2/3\\0&7/3&-7/3\end{pmatrix}\xrightarrow{L_3\leftarrow L_3+7L_2/2}\begin{pmatrix}-3&1&2\\0&-2/3&2/3\\0&0&0\end{pmatrix}$$

Le système admet une infinité de solutions de la forme (κ,κ,κ) avec $\kappa\in\mathbb{R}$.

Exercice 1.67
Équilibrer la réaction

$$Zn+H_2SO_4+K_2Cr_2O_7\rightarrow KCr(SO_4)_2+ZnSO_4+H_2O.$$

Correction
Écrivons les coefficients stœchiométriques et les contraintes :

$$x_1Zn+x_2H_2SO_4+x_3K_2Cr_2O_7\rightarrow x_4KCr(SO_4)_2+x_5ZnSO_4+x_6H_2O$$

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

$$\left(\begin{array}{ccccc|c}1&0&0&0&-1&0\\0&1&0&0&0&1\\0&1&0&-2&-1&0\\0&0&2&-1&0&0\\0&4&7&-8&-4&1\end{array}\right)\xrightarrow[L_5\leftarrow L_5-4L_2]{L_3\leftarrow L_3-L_2}\left(\begin{array}{ccccc|c}1&0&0&0&-1&0\\0&1&0&0&0&1\\0&0&0&-2&-1&-1\\0&0&2&-1&0&0\\0&0&7&-8&-4&-3\end{array}\right)\xrightarrow{L_3\leftrightarrow L_4}\left(\begin{array}{ccccc|c}1&0&0&0&-1&0\\0&1&0&0&0&1\\0&0&2&-1&0&0\\0&0&0&-2&-1&-1\\0&0&7&-8&-4&-3\end{array}\right)$$

$$\xrightarrow{L_5\leftarrow L_5-\frac{7}{2}L_3}\left(\begin{array}{ccccc|c}1&0&0&0&-1&0\\0&1&0&0&0&1\\0&0&2&-1&0&0\\0&0&0&-2&-1&-1\\0&0&0&-\frac{9}{2}&-4&-3\end{array}\right)\xrightarrow{L_5\leftarrow L_5-\frac{9}{4}L_4}\left(\begin{array}{ccccc|c}1&0&0&0&-1&0\\0&1&0&0&0&1\\0&0&2&-1&0&0\\0&0&0&-2&-1&-1\\0&0&0&0&-\frac{7}{4}&-\frac{3}{4}\end{array}\right)$$

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

$$3Zn + 7H_2SO_4 + K_2Cr_2O_7 \to 2KCr(SO_4)_2 + 3ZnSO_4 + 7H_2O.$$

Exercice 1.68 (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 :

$$\begin{cases} 7m+12a+6i=8, \\ 11m+6a+10i=9, \\ 11m+16a+14i=14, \end{cases} \xrightarrow[L_3 \leftarrow L_3 - \frac{11}{7}L_1]{L_2 \leftarrow L_2 - \frac{11}{7}L_1} \begin{cases} 7m+12a+6i=8, \\ -\frac{90}{7}a+\frac{4}{7}i=-\frac{25}{7}, \\ -\frac{20}{7}a+\frac{32}{7}i=\frac{10}{7}, \end{cases} \xrightarrow{L_3 \leftarrow L_3 - \frac{2}{9}L_2} \begin{cases} 7m+12a+6i=8, \\ -\frac{90}{7}a+\frac{4}{7}i=-\frac{25}{7}, \\ \frac{40}{9}i=\frac{20}{9}, \end{cases}$$

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 :

$$\begin{cases} -m+4a-2i=0, \\ 2m-3a+i=0, \\ -3m+2a=0, \end{cases} \xrightarrow[L_3 \leftarrow L_3 - \frac{11}{7}L_1]{L_2 \leftarrow L_2 - \frac{11}{7}L_1} \begin{cases} -m+4a-2i=0, \\ 5a-3i=0, \\ -10a+6i=0, \end{cases} \xrightarrow{L_3 \leftarrow L_3 - \frac{2}{9}L_2} \begin{cases} -m+4a-2i=0, \\ 5a-3i=0, \\ 0=0, \end{cases}$$

qui admet une infinité de solutions de la forme $(2\kappa, 3\kappa, 5\kappa)$ avec $\kappa \in [0;1/5]$.

Exercice 1.69 (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, ℓ 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

$$\mathbb{A}=\begin{pmatrix}0.4 & 0.5 & 0.6 & 0.7\\ 1 & 1.1 & 1.2 & 1.3\\ 1.5 & 1.7 & 1.9 & 2.1\end{pmatrix} \qquad \mathbb{B}=\begin{pmatrix}500 & 400 & 1000\\ 1000 & 900 & 700\\ 500 & 600 & 0\end{pmatrix} \qquad \mathbf{u}=\begin{pmatrix}s\\ m\\ \ell\\ x\end{pmatrix} \qquad \mathbf{v}=\begin{pmatrix}c\\ a\\ e\end{pmatrix}$$

1. Pour produire s manteaux taille S, m manteaux taille M, ℓ 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 1.70
Soit le système linéaire
$$(S)\quad\begin{cases}x-\alpha y=1,\\ \alpha x-y=1.\end{cases}$$
Déterminer les valeurs de α 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 1.71

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

$$\begin{cases}x+\alpha y=1\\-\alpha x-y=1\end{cases}\xrightarrow{L_2\leftarrow L_2+\alpha L_1}\begin{cases}x+\alpha y=1\\(-1+\alpha^2)y=\alpha+1\end{cases}$$

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 1.72

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 β de telle sorte que ce système possède :
1. une infinité de solutions ;
2. aucune solution ;
3. une solution unique.

Correction

$$\left(\begin{array}{ccc|c}1&1&-1&1\\2&3&\beta&3\\1&\beta&3&-3\end{array}\right)\xrightarrow[L_3\leftarrow L_3-L_1]{L_2\leftarrow L_2-2L_1}\left(\begin{array}{ccc|c}1&1&-1&1\\0&1&\beta+2&1\\0&\beta-1&4&-4\end{array}\right)\xrightarrow{L_3\leftarrow L_3+(1-\beta)L_2}\left(\begin{array}{ccc|c}1&1&-1&1\\0&1&\beta+2&1\\0&0&(6-\beta-\beta^2)&-(3+\beta)\end{array}\right).$$

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 1.73
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

$$\left(\begin{array}{ccc|c} 2 & -1 & 0 & \kappa \\ 1 & -1 & -1 & 0 \\ 1 & -\kappa & \kappa & \kappa \end{array}\right) \xrightarrow[L_3 \leftarrow L_3 - L_1/2]{L_2 \leftarrow L_2 - L_1/2} \left(\begin{array}{ccc|c} 2 & -1 & 0 & \kappa \\ 0 & -1/2 & -1 & -\kappa/2 \\ 0 & -\kappa+1/2 & \kappa & \kappa/2 \end{array}\right) \xrightarrow{L_3 \leftarrow L_3 + (1-2\kappa)L_2} \left(\begin{array}{ccc|c} 2 & -1 & 0 & \kappa \\ 0 & -1/2 & -1 & -\kappa/2 \\ 0 & 0 & 3\kappa-1 & \kappa^2 \end{array}\right).$$

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 κ pour que (S) possède une infinité de solutions.

Exercice 1.74
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

$$[\mathbb{A}|\mathbf{b}] = \left(\begin{array}{ccc|c} 1 & 2 & 3 & 2 \\ 1 & -1 & 2 & 7 \\ 3 & 0 & a & 10 \end{array}\right) \xrightarrow[L_3 \leftarrow L_3 - 3L_1]{L_2 \leftarrow L_2 - L_1} \left(\begin{array}{ccc|c} 1 & 2 & 3 & 2 \\ 0 & -3 & -1 & 5 \\ 0 & -6 & a-9 & 4 \end{array}\right) \xrightarrow{L_3 \leftarrow L_3 - 2L_2} \left(\begin{array}{ccc|c} 1 & 2 & 3 & 2 \\ 0 & -3 & -1 & 5 \\ 0 & 0 & a-7 & -6 \end{array}\right).$$

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 $a \neq 7$, $z = \dfrac{-6}{a-7}$, $y = \dfrac{5+z}{-3} = \dfrac{5a-41}{-3(a-7)}$ et $x = 2 - 2y - 3z = \dfrac{2(8a-35)}{3(a-7)}$ 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 $[\mathbb{A}|\mathbf{b}]$:

- $\operatorname{rg}(\mathbb{A}) = \begin{cases} 3 & \text{si } a \neq 7 \\ 2 & \text{si } a = 7 \end{cases}$ 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 ;
- $\operatorname{rg}([\mathbb{A}|\mathbf{b}]) = 3$ car $\det\begin{pmatrix} 1 & 2 & 2 \\ 1 & -1 & 7 \\ 3 & 0 & 10 \end{pmatrix} \neq 0$ 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 1.75
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

$$\left(\begin{array}{ccc|c} 1 & -1 & 2 & 7 \\ 1 & 2 & 3 & 2 \\ 3 & 0 & a & 10 \end{array}\right) \xrightarrow[L_3 \leftarrow L_3-3L_1]{L_2 \leftarrow L_2-L_1} \left(\begin{array}{ccc|c} 1 & -1 & 2 & 7 \\ 0 & 3 & 1 & -5 \\ 0 & 3 & a-6 & -11 \end{array}\right) \xrightarrow{L_3 \leftarrow L_3-L_2} \left(\begin{array}{ccc|c} 1 & -1 & 2 & 7 \\ 0 & 3 & 1 & -5 \\ 0 & 0 & a-7 & -6 \end{array}\right).$$

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 $a \neq 7$, $z=\dfrac{-6}{a-7}$, $y=\dfrac{-5-z}{3}=\dfrac{-5a+41}{3(a-7)}$ et $x=7-y-2z=\dfrac{2(8a-35)}{3(a-7)}$ 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 $[\mathbb{A}|\mathbf{b}]$:

- $\operatorname{rg}(\mathbb{A})=\begin{cases} 3 & \text{si } a \neq 7 \\ 2 & \text{si } a=7 \end{cases}$ 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 ;
- $\operatorname{rg}([\mathbb{A}|\mathbf{b}])=3$ car $\det\begin{pmatrix} 1 & 2 & 2 \\ 1 & -1 & 7 \\ 3 & 0 & 10 \end{pmatrix} \neq 0$ où $\begin{pmatrix} 1 & -1 & 2 \\ 1 & 2 & 7 \\ 3 & 0 & 10 \end{pmatrix}$ est la sous-matrice de $[\mathbb{A}|\mathbf{b}]A$ obtenue en supprimant la 3-ème colonne.

Exercice 1.76
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

$$\begin{pmatrix} 1 & 0 & 1 & 1 \\ a & 1 & a-1 & 1 \\ 2 & a & 1 & 2 \\ 1 & -1 & 2 & a \end{pmatrix} \xrightarrow[L_4 \leftarrow L_4-L_1]{\substack{L_2 \leftarrow L_2-aL_1 \\ L_3 \leftarrow L_3-2L_1}} \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & -1 & 1-a \\ 0 & a & -1 & 0 \\ 0 & -1 & 1 & a-1 \end{pmatrix} \xrightarrow[L_4 \leftarrow L_4+L_2]{L_3 \leftarrow L_3-aL_2} \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & -1 & 1-a \\ 0 & 0 & a-1 & a(a-1) \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

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 $\mathrm{Vect}\{(a-1,-1,-a,1)\}$ 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 $\mathrm{Vect}\{(-1,1,1,0),(-1,0,0,1)\}$ sont solution du système linéaire.

Exercice 1.77

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

$$\begin{pmatrix} 1 & 0 & 1 & 1 \\ b+1 & 1 & b & 1 \\ 2 & b+1 & 1 & 2 \\ 1 & -1 & 2 & b+1 \end{pmatrix} \xrightarrow[L_4 \leftarrow L_4 - L_1]{\substack{L_2 \leftarrow L_2 - (b+1)L_1 \\ L_3 \leftarrow L_3 - 2L_1}} \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & -1 & -b \\ 0 & b+1 & -1 & 0 \\ 0 & -1 & 1 & b \end{pmatrix} \xrightarrow[L_4 \leftarrow L_4 + L_2]{L_3 \leftarrow L_3 - (b+1)L_2} \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & -1 & -b \\ 0 & 0 & b & b(b+1) \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

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 $b \neq 0$,
$$z = \frac{-b(b+1)w}{b} = -(b+1)\kappa_1, \qquad y = bw + z = -\kappa_1, \qquad x = -w - z = b\kappa_1;$$
tous les vecteurs de $\mathrm{Vect}\{(b+1,1,b+1,-1)\}$ 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 $\mathrm{Vect}\{(-1,1,1,0),(-1,0,0,1)\}$ sont solution du système linéaire.

Exercice 1.78

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 :

$$\begin{vmatrix} 1+a & 1 & 1 \\ 1 & 1+a & 1 \\ 1 & 1 & 1+a \end{vmatrix} = (1+a)^3 + 1 + 1 - (1+a) - (1+a) - (1+a) = (1+a)^3 - 3(1+a) + 2$$
$$= \Big((1+a)-1\Big)\Big((1+a)^2 + (1+a) - 2\Big) = \Big((1+a)-1\Big)\Big((1+a)+2\Big)\Big((1+a)-1\Big) = a^2(3+a).$$

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 $\mathscr{S}$ 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 :

$$\begin{cases} -2x & y & +z=0 \\ x & -2y & +z=0 \\ x & y & -2z=0 \end{cases} \xrightarrow[L_3\leftarrow L_3+L_1/2]{L_2\leftarrow L_2+L_1/2} \begin{cases} -2x & y & +z=0 \\ & -\frac{3}{2}y & +\frac{3}{2}z=0 \\ & \frac{3}{2}y & -\frac{3}{2}z=0 \end{cases} \xrightarrow{L_3\leftarrow L_3+L_2} \begin{cases} -2x & -y & +z=0 \\ & -\frac{5}{2}y & +\frac{3}{2}z=0 \\ & & 0z=0 \end{cases}$$

donc $z=\kappa\in\mathbb{R}$, $y=z$ et $x=z$, ainsi

$$\mathscr{S}=\{(\kappa,\kappa,\kappa)\mid \kappa\in\mathbb{R}\}.$$

⋆ *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

$$\mathscr{S}=\left\{(-\kappa_1-\kappa_2,\kappa_2,\kappa_1)\;\middle|\;(\kappa_1,\kappa_2)\in\mathbb{R}^2\right\}.$$

⋆ *Étude du cas* $a\in\mathbb{R}\setminus\{-3,0\}$. Il s'agit d'un système de Cramer homogène, donc l'unique solution est $(0,0,0)$:

$$\mathscr{S}=\{(0,0,0)\}.$$

Exercice 1.79
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 $\mathscr{S}$ 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

$$\mathscr{S}=\left\{\left(\kappa,\frac{3-3\kappa}{2},\kappa\right)\;\middle|\;\kappa\in\mathbb{R}\right\}.$$

⋆ *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 :

$$\begin{cases} x+4y+3z=0,\\ 3x+2y+4z=3,\\ 3x+4y+5z=4,\end{cases} \xrightarrow[L_3\leftarrow L_3-3L_1]{L_2\leftarrow L_2-3L_1} \begin{cases} x\ +4y+3z=0,\\ -10y-5z=3,\\ -8y-4z=4,\end{cases} \xrightarrow{L_3\leftarrow 10L_3-8L_2} \begin{cases} x\ +4y+3z=0,\\ -10y-5z=3,\\ 0=16.\end{cases}$$

La dernière équation est impossible donc

$$\mathscr{S}=\emptyset.$$

⋆ *Étude du cas* $a\in\mathbb{R}\setminus\{-3,0\}$. On utilise la méthode du pivot de GAUSS :

$$\begin{cases} x+ay+(a-1)z=0,\\ 3x+2y\ \ \ +az=3,\\ (a-1)x+ay+(a+1)z=a,\end{cases} \xrightarrow[L_3\leftarrow L_3-(a-1)L_1]{L_2\leftarrow L_2-3L_1} \begin{cases} x\ \ \ +ay\ \ +(a-1)z=0,\\ (2-3a)y+(3-2a)z=3,\\ (2-a)ay+(3-a)az=a,\end{cases}$$

$$\xrightarrow{L_3\leftarrow L_3-\frac{(2-a)a}{(2-3a)}L_2} \begin{cases} x\ \ \ +ay\ \ +(a-1)z=0,\\ (2-3a)y+(3-2a)z=3,\\ -\frac{a^2(a-4)}{3a-2}z=\frac{4a}{3a-2}.\end{cases}$$

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

$$\mathscr{S}=\left\{\left(\frac{a^2-2a-4}{a(a-4)},-\frac{a-6}{a(a-4)},-\frac{4}{a(a-4)}\right)\right\}.$$

Calcul de la matrice inverse

Exercice 1.80
Calculer $\mathbb{A}^{-1}$ où $\mathbb{A}$ est la matrice $\begin{pmatrix}1&0&-1\\4&-1&-2\\-2&0&1\end{pmatrix}$.

Correction

$$[\mathbb{A}|\mathbb{I}_3]=\left(\begin{array}{ccc|ccc}1&0&-1&1&0&0\\4&-1&-2&0&1&0\\-2&0&1&0&0&1\end{array}\right) \xrightarrow[L_3\leftarrow L_3+2L_1]{\substack{L_1\leftarrow L_1\\ L_2\leftarrow L_2-4L_1}} \left(\begin{array}{ccc|ccc}1&0&-1&1&0&0\\0&-1&2&-4&1&0\\0&0&-1&-2&0&1\end{array}\right)$$

$$\xrightarrow[L_3\leftarrow L_3]{\substack{L_1\leftarrow L_1\\ L_2\leftarrow -L_2}} \left(\begin{array}{ccc|ccc}1&0&-1&1&0&0\\0&1&-2&4&-1&0\\0&0&-1&-2&0&1\end{array}\right) \xrightarrow[L_3\leftarrow -L_3]{\substack{L_1\leftarrow L_1-L_3\\ L_2\leftarrow L_2-2L_3}} \left(\begin{array}{ccc|ccc}1&0&0&-1&0&-1\\0&1&0&0&-1&-2\\0&0&1&-2&0&-1\end{array}\right)=[\mathbb{I}_3|\mathbb{A}^{-1}].$$

```
A=[1 0 -1; 4 -1 -2; -2 0 1]
inv(A)
```

Exercice 1.81
Calculer $\mathbb{A}^{-1}$ où $\mathbb{A}$ est la matrice $\begin{pmatrix}1&0&1\\0&1&2\\2&0&1\end{pmatrix}$.

Correction

$$[\mathbb{A}|\mathbb{I}_3] = \left(\begin{array}{ccc|ccc} 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 2 & 0 & 1 & 0 \\ 2 & 0 & 1 & 0 & 0 & 1 \end{array}\right) \xrightarrow[L_3 \leftarrow L_3 - 2L_1]{\substack{L_1 \leftarrow L_1 \\ L_2 \leftarrow L_2}} \left(\begin{array}{ccc|ccc} 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 2 & 0 & 1 & 0 \\ 0 & 0 & -1 & -2 & 0 & 1 \end{array}\right)$$

$$\xrightarrow[L_3 \leftarrow L_3]{\substack{L_1 \leftarrow L_1 \\ L_2 \leftarrow L_2}} \left(\begin{array}{ccc|ccc} 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 2 & 0 & 1 & 0 \\ 0 & 0 & -1 & -2 & 0 & 1 \end{array}\right) \xrightarrow[L_3 \leftarrow -L_3]{\substack{L_1 \leftarrow L_1 + L_3 \\ L_2 \leftarrow L_2 + 2L_3}} \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & -1 & 0 & 1 \\ 0 & 1 & 0 & -4 & 1 & 2 \\ 0 & 0 & 1 & 2 & 0 & -1 \end{array}\right) = [\mathbb{I}_3|\mathbb{A}^{-1}].$$

```
A=[1 0 1; 0 1 2; 2 0 1]
inv(A)
```

Exercice 1.82

Calculer les inverses des matrices suivantes (si elles existent) :

$$\mathbb{A} = \begin{pmatrix} 2 & -3 \\ 4 & 5 \end{pmatrix}, \qquad \mathbb{B} = \begin{pmatrix} 1 & 5 & -3 \\ 2 & 11 & 1 \\ 2 & 9 & -11 \end{pmatrix}, \qquad \mathbb{C} = \begin{pmatrix} 1 & 5 & -3 \\ 2 & 11 & 1 \\ 1 & 4 & -10 \end{pmatrix}.$$

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 1.83

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

$$[\mathbb{A}|\mathbb{I}_4] = \left(\begin{array}{cccc|cccc} 1&0&0&-1&1&0&0&0\\ 1&1&-1&-1&0&1&0&0\\ 1&2&-1&-2&0&0&1&0\\ 1&2&0&-2&0&0&0&1 \end{array}\right) \xrightarrow[L_4\leftarrow L_4-L_1]{\substack{L_2\leftarrow L_2-L_1\\ L_3\leftarrow L_3-L_1}} \left(\begin{array}{cccc|cccc} 1&0&0&-1&1&0&0&0\\ 0&1&-1&0&-1&1&0&0\\ 0&2&-1&-1&-1&0&1&0\\ 0&2&0&-1&-1&0&0&1 \end{array}\right)$$

$$\xrightarrow[L_4\leftarrow L_4-2L_2]{\substack{L_1\leftarrow L_1\\ L_3\leftarrow L_3-2L_2}} \left(\begin{array}{cccc|cccc} 1&0&0&-1&1&0&0&0\\ 0&1&-1&0&-1&1&0&0\\ 0&0&1&-1&1&-2&1&0\\ 0&0&2&-1&1&-2&0&1 \end{array}\right) \xrightarrow[L_4\leftarrow L_4-2L_3]{\substack{L_1\leftarrow L_1\\ L_2\leftarrow L_2+L_3}} \left(\begin{array}{cccc|cccc} 1&0&0&-1&1&0&0&0\\ 0&1&0&-1&0&-1&1&0\\ 0&0&1&-1&1&-2&1&0\\ 0&0&0&1&-1&2&-2&1 \end{array}\right)$$

$$\xrightarrow[L_3\leftarrow L_3+L_4]{\substack{L_1\leftarrow L_1+L_4\\ L_2\leftarrow L_2+L_4}} \left(\begin{array}{cccc|cccc} 1&0&0&0&0&2&-2&1\\ 0&1&0&0&-1&1&-1&1\\ 0&0&1&0&0&0&-1&1\\ 0&0&0&1&-1&2&-2&1 \end{array}\right) = [\mathbb{I}_4|\mathbb{A}^{-1}].$$

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)
```

Systèmes linéaires : méthode de Gauss pour des systèmes rectangulaires (sur ou sous déterminés)

Exercice 1.84

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 forcement des solutions ; comme il y a 4 inconnues, alors on a une infinité de solutions.

Exercice 1.85
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 (κ,κ,κ) pour $\kappa \in \mathbb{R}$.

Exercice 1.86
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 ∞^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 1.87
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

$$\begin{cases} x+2y+z=-1, \\ 2x+y-z=1, \\ -x+y+2z=-2, \end{cases} \xrightarrow[L_3 \leftarrow L_3+L_1]{L_2 \leftarrow L_2-2L_1} \begin{cases} x+2y+z=-1, \\ -3y-3z=3, \\ 3y+3z=-3, \end{cases} \xrightarrow{L_3 \leftarrow L_3+L_2} \begin{cases} x+2y+z=-1, \\ -3y-3z=3, \\ 0=0, \end{cases}$$

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 1.88
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

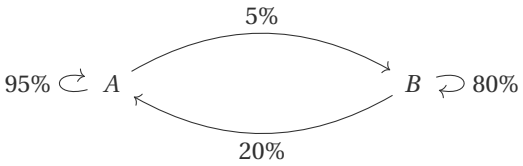
$$\begin{cases} x-2y+2z=0, \\ 2x+y-2z=0, \\ 3x+4y-6z=0, \end{cases} \xrightarrow[L_3\leftarrow L_3-3L_1]{L_2\leftarrow L_2-2L_1} \begin{cases} x-2y+2z=0, \\ 5y-6z=0, \\ 10y-12z=0, \end{cases} \xrightarrow{L_3\leftarrow L_3-2L_2} \begin{cases} x-2y+2z=0, \\ 5y-6z=0, \\ 0=0, \end{cases}$$

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 $\kappa \in \mathbb{R}$.

1.5.4. Valeurs et vecteur propres

Exercice 1.89 (Migration entre deux villes)
Deux villes A et B totalisent une population d'un million d'habitants. La ville A est plus agréable, mais la ville B offre plus de possibilités d'emplois. 20% des habitants de B partent chaque année habiter A pour avoir un meilleur cadre de vie, et 5% des habitants de A partent chaque année habiter B pour trouver un meilleur emploi. Sachant qu'à l'année 0, un quart des habitants sont en A, quelle est la population de A et de B au bout de 1 an, 2 ans, 4 ans, 9 ans?

Correction
On résume les informations dans un graphe de transition :



Les sommets du graphes correspondent aux différents états possibles (ici, habiter la ville A ou la ville B), et les flèches donnent le pourcentage de gens qui passent d'un état à un autre, d'une année sur l'autre.

Méthode directe La suite des états successifs est décrite par une relation de récurrence linéaire, de la forme $\mathbf{x}_{n+1} = \mathbb{T}\mathbf{x}_n$. Le vecteur $\mathbf{x}_n \in \mathbb{R}^2$ est le vecteur d'état du système, *i.e.* le vecteur (a_n, b_n) où a_n est la population de la ville A après n années et b_n est la population de la ville B après n années, et la matrice $\mathbb{T}$ est la matrice de transition. Ainsi,

$$\mathbf{x}^{(0)} = \begin{pmatrix} \frac{1}{4} \\ \\ \frac{3}{4} \end{pmatrix} \qquad \mathbb{T} = \begin{pmatrix} 95\% & 20\% \\ \\ 5\% & 80\% \end{pmatrix}$$

donc

$$\mathbf{x}^{(1)} = \mathbb{T}\mathbf{x}^{(0)} = \begin{pmatrix} 95\% & 20\% \\ \\ 5\% & 80\% \end{pmatrix} \begin{pmatrix} \frac{1}{4} \\ \\ \frac{3}{4} \end{pmatrix} = \begin{pmatrix} \frac{31}{80} \\ \\ \frac{49}{80} \end{pmatrix} = \begin{pmatrix} 0.3875 \\ \\ 0.6125 \end{pmatrix}$$

$$\mathbf{x}^{(2)} = \mathbb{T}\mathbf{x}^{(1)} = \begin{pmatrix} 95\% & 20\% \\ 5\% & 80\% \end{pmatrix} \begin{pmatrix} \frac{31}{80} \\ \frac{49}{80} \end{pmatrix} = \begin{pmatrix} \frac{157}{320} \\ \frac{163}{320} \end{pmatrix} = \begin{pmatrix} 0.90625 \\ 0.509375 \end{pmatrix}$$

Méthode par récurrence La relation de récurrence linéaire qui peut être explicitée : $\mathbf{x}_n = \mathbb{T}^n\mathbf{x}_0$. Ainsi,

$$\mathbf{x}^{(2)} = \mathbb{T}^2\mathbf{x}^{(0)} = \begin{pmatrix} \frac{73}{80} & \frac{7}{20} \\ \frac{7}{80} & \frac{13}{20} \end{pmatrix} \begin{pmatrix} \frac{1}{4} \\ \frac{3}{4} \end{pmatrix} = \begin{pmatrix} \frac{157}{320} \\ \frac{163}{320} \end{pmatrix} = \begin{pmatrix} 0.90625 \\ 0.509375 \end{pmatrix}$$
$$\mathbf{x}^{(4)} = \mathbb{T}^4\mathbf{x}^{(0)} = \begin{pmatrix} \frac{221}{256} & \frac{35}{64} \\ \frac{35}{256} & \frac{29}{64} \end{pmatrix} \begin{pmatrix} \frac{1}{4} \\ \frac{3}{4} \end{pmatrix} = \begin{pmatrix} \frac{641}{1024} \\ \frac{383}{1024} \end{pmatrix} \simeq \begin{pmatrix} 0.6259765625 \\ 0.3740234375 \end{pmatrix}$$
$$\mathbf{x}^{(9)} = \mathbb{T}^9\mathbf{x}^{(0)} = \begin{pmatrix} \frac{1068259}{1310720} & \frac{242461}{327680} \\ \frac{242461}{1310720} & \frac{85219}{327680} \end{pmatrix} \begin{pmatrix} \frac{1}{4} \\ \frac{3}{4} \end{pmatrix} = \begin{pmatrix} \frac{3977791}{5242880} \\ \frac{1265089}{5242880} \end{pmatrix} \simeq \begin{pmatrix} 0.7587034225 \\ 0.2412965775 \end{pmatrix}$$

On peut vérifier numériquement que

$$\lim_{n\to+\infty} \mathbf{x}^{(n)} = \begin{pmatrix} 0.8 \\ 0.2 \end{pmatrix}$$

autrement dit la suite converge vers un état où le 80% de la population se trouve dans la la ville A et 20% dans la ville B.

Méthode par diagonalisation En diagonalisant d'abord la matrice $\mathbb{T}$ on obtient $\mathbb{T} = \mathbb{P}\mathbb{D}\mathbb{P}^{-1}$ avec

$$\mathbb{P} = \begin{pmatrix} \frac{4}{\sqrt{17}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{17}} & \frac{1}{\sqrt{2}} \end{pmatrix} \qquad \mathbb{D} = \begin{pmatrix} 1 & 0 \\ 0 & 25\% \end{pmatrix} \qquad \mathbb{P}^{-1} = \begin{pmatrix} \frac{\sqrt{17}}{5} & \frac{\sqrt{17}}{5} \\ -\frac{\sqrt{2}}{5} & \frac{4\sqrt{2}}{5} \end{pmatrix}$$

donc

$$\mathbf{x}^{(k)} = \mathbb{P}\mathbb{D}^k\mathbb{P}^{-1}\mathbf{x}^{(0)} = \begin{pmatrix} \frac{4}{\sqrt{17}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{17}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{4^k} \end{pmatrix} \begin{pmatrix} \frac{\sqrt{17}}{5} & \frac{\sqrt{17}}{5} \\ -\frac{\sqrt{2}}{5} & \frac{4\sqrt{2}}{5} \end{pmatrix} \begin{pmatrix} \frac{1}{4} \\ \frac{3}{4} \end{pmatrix}$$
$$\xrightarrow[k\to+\infty]{} \begin{pmatrix} \frac{4}{\sqrt{17}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{17}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{\sqrt{17}}{5} & \frac{\sqrt{17}}{5} \\ -\frac{\sqrt{2}}{5} & \frac{4\sqrt{2}}{5} \end{pmatrix} \begin{pmatrix} \frac{1}{4} \\ \frac{3}{4} \end{pmatrix} = \begin{pmatrix} \frac{4}{5} & \frac{4}{5} \\ \frac{1}{5} & \frac{1}{5} \end{pmatrix} \begin{pmatrix} \frac{1}{4} \\ \frac{3}{4} \end{pmatrix} = \begin{pmatrix} \frac{4}{5} \\ \frac{1}{5} \end{pmatrix}$$

Exercice 1.90 (Chaîne de Markov)

Considérons un processus de MARKOV modélisé par la matrice de transition suivante :

$$\mathbb{T} = \begin{pmatrix} 0 & \frac{1}{3} \\ 1 & \frac{2}{3} \end{pmatrix}$$

1. Vérifier que les valeurs propres de $\mathbb{T}$ sont 1 et $-\frac{1}{3}$.
2. Calculer les vecteurs propres associés à ces valeurs propres sans les normaliser (on utilisera la méthode de GAUSS pour résoudre les deux systèmes linéaires).
3. Définir deux matrice $\mathbb{D}$ et $\mathbb{P}$ telles que $\mathbb{T} = \mathbb{P}\mathbb{D}\mathbb{P}^{-1}$ et calculer $\mathbb{P}^{-1}$ en utilisant la méthode de GAUSS.
4. On veut trouver le comportement du processus à long terme si l'état initiale est $\mathbf{x}^{(0)} = (1,0)^T$. On sait que le comportement du processus à l'étape $k+1$ est lié au comportement à l'étape k par la relation $\mathbf{x}^{(k+1)} = \mathbb{T}\mathbf{x}^{(k)}$ et donc, par récurrence, $\mathbf{x}^{(k+1)} = \mathbb{T}^{k+1}\mathbf{x}^{(0)}$. On cherche à calculer $\lim_{k\to+\infty} \mathbf{x}^{(k)}$.
   - ⋆ S'il existe $\mathbf{x} = \lim_{k\to+\infty} \mathbf{x}^{(k)}$, alors $\mathbf{x} = \mathbb{T}\mathbf{x}$, autrement dit $\mathbf{x}$ est solution du système linéaire $(\mathbb{T} - \mathbb{I})\mathbf{x} = \mathbf{0}$.
     Calculer $\mathbf{x}$ et le diviser par la somme de ses composantes.
   - ⋆ Étant donné que $\mathbb{T} = \mathbb{P}\mathbb{D}\mathbb{P}^{-1}$, alors $\lim_{k\to+\infty} \mathbf{x}^{(k)} = \mathbb{P}(\lim_{k\to+\infty} \mathbb{D}^k)\mathbb{P}^{-1}\mathbf{x}^{(0)}$.
     Calculer le produit $\mathbb{P}(\lim_{k\to+\infty} \mathbb{D}^k)\mathbb{P}^{-1}\mathbf{x}^{(0)}$ et le diviser par la somme de ses composantes.
     Vérifier qu'on obtient bien la valeur calculée précédemment.

Correction

1. Calcul des valeurs propres :

$$p(\lambda) \overset{\text{def}}{=} \det(\mathbb{T} - \lambda \mathbb{I}_2) = \det\begin{pmatrix} -\lambda & 1 \\ \frac{1}{3} & \frac{2}{3} - \lambda \end{pmatrix} = -\lambda\left(\frac{2}{3} - \lambda\right) - \frac{1}{3} = \lambda^2 - \frac{2}{3}\lambda - \frac{1}{3}.$$

On vérifie facilement que les valeurs données annulent ce polynôme, en effet :

$$p(1) = 1^2 - \frac{2}{3} - \frac{1}{3} = 0,$$
$$p\left(-\frac{1}{3}\right) = \left(-\frac{1}{3}\right)^2 + \frac{2}{3}\frac{1}{3} - \frac{1}{3} = \frac{1}{9} + \frac{2}{9} - \frac{1}{3} = 0.$$

2. On pose $\lambda_1 = 1$ et $\lambda_2 = -\frac{1}{3}$. Pour calculer les vecteurs propres on doit résoudre deux systèmes linéaires homogènes. [1]

2.1. On résout le système linéaire $(\mathbb{T} - \lambda_1 \mathbb{I}_2)\mathbf{x} = \mathbf{0}$, ce qui donne l'espace vectoriel des vecteurs de la forme $(\kappa, 3\kappa)^T$:

$$\begin{pmatrix} -1 & \frac{1}{3} \\ 1 & -\frac{1}{3} \end{pmatrix} \xrightarrow[L_2 \leftarrow L_2 + \frac{1}{3}L_1]{} \begin{pmatrix} -1 & \frac{1}{3} \\ 0 & 0 \end{pmatrix} \text{ donc } \begin{cases} y = \kappa, \\ x = \frac{1}{3}y = \frac{\kappa}{3}. \end{cases}$$

On choisit par exemple comme base l'élément $\mathbf{x} = (1,3)^T$.

2.2. On résout le système linéaire $(\mathbb{T} - \lambda_2 \mathbb{I}_2)\mathbf{x} = \mathbf{0}$, ce qui donne l'espace vectoriel des vecteurs de la forme $(-\kappa, \kappa)^T$:

$$\begin{pmatrix} \frac{1}{3} & \frac{1}{3} \\ 1 & 1 \end{pmatrix} \xrightarrow[L_2 \leftarrow L_2 - L_1]{} \begin{pmatrix} \frac{1}{3} & \frac{1}{3} \\ 0 & 0 \end{pmatrix} \text{ donc } \begin{cases} y = \kappa, \\ x = -y = -\kappa. \end{cases}$$

On choisit par exemple comme base l'élément $\mathbf{x} = (-1,1)^T$.

3. $\mathbb{D}$ est la matrice diagonale contenant les valeurs propres et $\mathbb{P}$ la matrice dont chaque colonne contient un vecteur de l'espace propre associé. On pose donc

$$\mathbb{D} = \begin{pmatrix} 1 & 0 \\ 0 & -\frac{1}{3} \end{pmatrix} \qquad \mathbb{P} = \begin{pmatrix} 1 & -1 \\ 3 & 1 \end{pmatrix}$$

et on calcule $\mathbb{P}^{-1}$:

$$[\mathbb{P}|\mathbb{I}_2] = \left(\begin{array}{cc|cc} 1 & -1 & 1 & 0 \\ 3 & 1 & 0 & 1 \end{array}\right) \xrightarrow{L_2 \leftarrow L_2 - 3L_1} \left(\begin{array}{cc|cc} 1 & -1 & 1 & 0 \\ 0 & 4 & -3 & 1 \end{array}\right)$$
$$\xrightarrow{L_2 \leftarrow L_2/4} \left(\begin{array}{cc|cc} 1 & -1 & 1 & 0 \\ 0 & 1 & -3/4 & 1/4 \end{array}\right) \xrightarrow{L_1 \leftarrow L_1 + L_2} \left(\begin{array}{cc|cc} 1 & 0 & 1/4 & 1/4 \\ 0 & 1 & -3/4 & 1/4 \end{array}\right) = [\mathbb{I}_2|\mathbb{P}^{-1}]$$

Ainsi

$$\mathbb{P}^{-1} = \begin{pmatrix} \frac{1}{4} & \frac{1}{4} \\ -\frac{3}{4} & \frac{1}{4} \end{pmatrix}$$

4. À long terme le vecteur d'état sera $\mathbf{x} = \lim_{k \to +\infty} \mathbf{x}^{(k)}$.

⋆ On a déjà calculé la solution du système linéaire $(\mathbb{T} - \mathbb{I}_2)\mathbf{x} = \mathbf{0}$, il s'agit simplement d'un vecteur propre associé à la valeur propre 1, donc si on note $\mathbf{v} = (1,3)^T$, on calcule $\mathbf{x} = \frac{\mathbf{v}}{\sum_{i=1}^2 v_i} = (1/4, 3/4)^T = (25\%, 75\%)^T$.

⋆ Étant donné que $\mathbb{T} = \mathbb{P}\mathbb{D}\mathbb{P}^{-1}$, alors

$$\mathbb{P}(\lim_{k \to +\infty} \mathbb{D}^k)\mathbb{P}^{-1}\mathbf{x}^{(0)} = \begin{pmatrix} 1 & -1 \\ 3 & 1 \end{pmatrix} \lim_{k \to +\infty} \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{(-3)^k} \end{pmatrix} \begin{pmatrix} \frac{1}{4} & \frac{1}{4} \\ -\frac{3}{4} & \frac{1}{4} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 3 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{4} & \frac{1}{4} \\ -\frac{3}{4} & \frac{1}{4} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$
$$= \begin{pmatrix} 1 & -1 \\ 3 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{4} & \frac{1}{4} \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{4} & \frac{1}{4} \\ \frac{3}{4} & \frac{3}{4} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{4} \\ \frac{3}{4} \end{pmatrix}$$

donc $\mathbf{x} = \left(\frac{1}{4}, \frac{3}{4}\right)^T$.

1. Les systèmes étant homogènes, on n'écrira pas la matrice augmentée.

Exercice 1.91 (Chaîne de Markov)

Une étude non officielle de la météo dans une ville au début du printemps montre les observations suivantes :

- ⋆ il est presque impossible d'avoir deux beaux jours consécutifs,
- ⋆ si nous avons un beau jour, on a la même probabilité d'avoir de la neige ou de la pluie le jour suivant,
- ⋆ si nous avons la neige ou de la pluie, alors nous avons une chance égale pour avoir la même chose le jour suivant,
- ⋆ s'il y a un changement de neige ou de pluie, seulement la moitié du temps ce changement est à un beau jour.

Si les lettres b, p, n représentent beau, pluie et neige respectivement, on note $P(i \to j)$ la probabilité d'avoir la météo j si la veille la météo était i, donc

$$\begin{aligned} &P(b\to b)=0 && P(p\to b)=0.25 && P(n\to b)=0.25\\ &P(b\to p)=0.5 && P(p\to p)=0.5 && P(n\to p)=0.25\\ &P(b\to n)=0.5 && P(p\to n)=0.25 && P(n\to n)=0.5. \end{aligned}$$

Comme la météo de demain dépend seulement d'aujourd'hui, c'est un processus de MARKOV. La matrice de transition qui modélise ce système est donc

$$\mathbb{T}=\begin{pmatrix} 0 & \frac{1}{4} & \frac{1}{4}\\ \frac{1}{2} & \frac{1}{2} & \frac{1}{4}\\ \frac{1}{2} & \frac{1}{4} & \frac{1}{2}\end{pmatrix}$$

1. Vérifier que les valeurs propres de $\mathbb{T}$ sont 1, $\frac{1}{4}$ et $-\frac{1}{4}$.
2. Calculer les vecteurs propres associés à ces valeurs propres sans les normaliser (on utilisera la méthode de GAUSS pour résoudre les trois systèmes linéaires).
3. Définir deux matrice $\mathbb{D}$ et $\mathbb{P}$ telles que $\mathbb{T}=\mathbb{P}\mathbb{D}\mathbb{P}^{-1}$ et calculer $\mathbb{P}^{-1}$ en utilisant la méthode de GAUSS.
4. On veut trouver le comportement de la météo à long terme s'il fait beau aujourd'hui, *i.e.* si $\mathbf{x}^{(0)}=(1,0,0)^T$. On sait que le comportement de la météo au jour $k+1$ est lié à la météo au jour k par la relation

   $$\mathbf{x}^{(k+1)}=\mathbb{T}\mathbf{x}^{(k)}$$

   et donc, par récurrence,

   $$\mathbf{x}^{(k+1)}=\mathbb{T}^{k+1}\mathbf{x}^{(0)}.$$

   On cherche à calculer $\lim\limits_{k\to+\infty}\mathbf{x}^{(k)}$.

   - ⋆ S'il existe $\mathbf{x}=\lim\limits_{k\to+\infty}\mathbf{x}^{(k)}$, alors $\mathbf{x}=\mathbb{T}\mathbf{x}$, autrement dit $\mathbf{x}$ est solution du système linéaire $(\mathbb{T}-\mathbb{I})\mathbf{x}=\mathbf{0}$. Calculer cette limite. [a]
   - ⋆ Étant donné que $\mathbb{T}=\mathbb{P}\mathbb{D}\mathbb{P}^{-1}$, alors

     $$\lim_{k\to+\infty}\mathbf{x}^{(k)}=\lim_{k\to+\infty}\mathbf{x}^{(k)}=\lim_{k\to+\infty}\mathbb{T}^k\mathbf{x}^{(0)}=\lim_{k\to+\infty}(\mathbb{P}\mathbb{D}\mathbb{P}^{-1})^k\mathbf{x}^{(0)}=\lim_{k\to+\infty}\mathbb{P}\mathbb{D}^k\mathbb{P}^{-1}\mathbf{x}^{(0)}=\mathbb{P}(\lim_{k\to+\infty}\mathbb{D}^k)\mathbb{P}^{-1}\mathbf{x}^{(0)}$$

     Calculer ce produit et vérifier qu'on obtient bien la limite $\mathbf{x}$ calculée précédemment.

a. Puisque toutes les entrées de la matrice de transition sont entre 0 et 1 exclusivement, alors la convergence est garantie d'avoir lieu. La convergence peut avoir lieu quand 0 et 1 sont dans la matrice de transition, mais la convergence n'est plus garantie.

Correction

1. Calcul des valeurs propres :

   $$p(\lambda)\overset{\text{def}}{=}\det(\mathbb{T}-\lambda\mathbb{I})=-\lambda\left(\frac{1}{2}-\lambda\right)^2+\frac{1}{32}+\frac{1}{32}-\frac{1}{8}\left(\frac{1}{2}-\lambda\right)+\frac{1}{16}\lambda-\frac{1}{8}\left(\frac{1}{2}-\lambda\right)=-\lambda^3+\lambda^2+\frac{1}{16}\lambda-\frac{1}{16}.$$

   On vérifie facilement que les valeurs données annulent ce polynôme, en effet :

   $$\begin{aligned} p(1)&=-\lambda^3+\lambda^2+\frac{1}{16}\lambda-\frac{1}{16}=-1+1+\frac{1}{16}-\frac{1}{16}=0,\\ p\left(\frac{1}{4}\right)&=-\frac{1}{4^3}+\frac{1}{4^2}+\frac{1}{16}\frac{1}{4}-\frac{1}{16}=0,\\ p\left(-\frac{1}{4}\right)&=-\frac{1}{(-4)^3}+\frac{1}{(-4)^2}+\frac{1}{16}\frac{1}{-4}-\frac{1}{16}=0. \end{aligned}$$

2. On pose $\lambda_1 = 1$, $\lambda_2 = \frac{1}{4}$ et $\lambda_3 = -\frac{1}{4}$. Pour calculer les vecteurs propres on doit résoudre trois systèmes linéaires. [2]

   2.1. On résout le système linéaire $(\mathbb{T} - \lambda_1 \mathbb{I})\mathbf{x} = \mathbf{0}$, ce qui donne $\mathbf{x} = (1,2,2)^T$:

$$\begin{pmatrix} -1 & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{4} & -\frac{1}{2} \end{pmatrix} \xrightarrow[L_3 \leftarrow L_3 + \frac{1}{2}L_1]{L_2 \leftarrow L_2 + \frac{1}{2}L_1} \begin{pmatrix} -1 & \frac{1}{4} & \frac{1}{4} \\ 0 & -\frac{3}{8} & \frac{3}{8} \\ 0 & \frac{3}{8} & -\frac{3}{8} \end{pmatrix} \xrightarrow{L_3 \leftarrow L_3 + L_2} \begin{pmatrix} -1 & \frac{1}{4} & \frac{1}{4} \\ 0 & -\frac{3}{8} & \frac{3}{8} \\ 0 & 0 & 0 \end{pmatrix} \text{ donc } \begin{cases} z = \kappa, \\ y = \frac{-\frac{3}{8}z}{-\frac{3}{8}} = \kappa, \\ x = \frac{-\frac{1}{4}y - \frac{1}{4}z}{-1} = \frac{\kappa}{2}. \end{cases}$$

   2.2. On résout le système linéaire $(\mathbb{T} - \lambda_2 \mathbb{I})\mathbf{x} = \mathbf{0}$, ce qui donne $\mathbf{x} = (0,1,-1)^T$.

$$\begin{pmatrix} -\frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{pmatrix} \xrightarrow[L_3 \leftarrow L_3 + 2L_1]{L_2 \leftarrow L_2 + 2L_1} \begin{pmatrix} -\frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{3}{4} & \frac{3}{4} \\ 0 & \frac{3}{4} & \frac{3}{4} \end{pmatrix} \xrightarrow{L_3 \leftarrow L_3 - L_2} \begin{pmatrix} -\frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{3}{4} & \frac{3}{4} \\ 0 & 0 & 0 \end{pmatrix} \text{ donc } \begin{cases} z = \kappa, \\ y = \frac{-\frac{3}{4}z}{\frac{3}{4}} = -\kappa, \\ x = \frac{-\frac{1}{4}y - \frac{1}{4}z}{-\frac{1}{4}} = 0. \end{cases}$$

   2.3. On résout le système linéaire $(\mathbb{T} - \lambda_3 \mathbb{I})\mathbf{x} = \mathbf{0}$, ce qui donne $\mathbf{x} = (-2,1,1)^T$.

$$\begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{3}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{4} & \frac{3}{4} \end{pmatrix} \xrightarrow[L_3 \leftarrow L_3 - 2L_1]{L_2 \leftarrow L_2 - 2L_1} \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{4} & -\frac{1}{4} \\ 0 & -\frac{1}{4} & \frac{1}{4} \end{pmatrix} \xrightarrow{L_3 \leftarrow L_3 + L_2} \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{4} & -\frac{1}{4} \\ 0 & 0 & 0 \end{pmatrix} \text{ donc } \begin{cases} z = \kappa, \\ y = \frac{\frac{1}{4}z}{\frac{1}{4}} = \kappa, \\ x = \frac{-\frac{1}{4}y - \frac{1}{4}z}{\frac{1}{4}} = -2\kappa. \end{cases}$$

3. $\mathbb{D}$ est la matrice diagonale contenant les valeurs propres et $\mathbb{P}$ la matrice dont chaque colonne contient le vecteur propre associé. On pose donc

$$\mathbb{D} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ 0 & 0 & -\frac{1}{4} \end{pmatrix} \qquad \mathbb{P} = \begin{pmatrix} 1 & 0 & -2 \\ 2 & 1 & 1 \\ 2 & -1 & 1 \end{pmatrix}$$

   et on calcule $\mathbb{P}^{-1}$:

$$[\mathbb{P}|\mathbb{I}_3] = \left(\begin{array}{ccc|ccc} 1 & 0 & -2 & 1 & 0 & 0 \\ 2 & 1 & 1 & 0 & 1 & 0 \\ 2 & -1 & 1 & 0 & 0 & 1 \end{array}\right) \xrightarrow[L_3 \leftarrow L_3 - 2L_1]{L_2 \leftarrow L_2 - 2L_1} \left(\begin{array}{ccc|ccc} 1 & 0 & -2 & 1 & 0 & 0 \\ 0 & 1 & 5 & -2 & 1 & 0 \\ 0 & -1 & 5 & -2 & 0 & 1 \end{array}\right)$$

$$\xrightarrow[L_3 \leftarrow L_3 + L_2]{L_1 \leftarrow L_1} \left(\begin{array}{ccc|ccc} 1 & 0 & -2 & 1 & 0 & 0 \\ 0 & 1 & 5 & -2 & 1 & 0 \\ 0 & 0 & 10 & -4 & 1 & 1 \end{array}\right) \xrightarrow[L_2 \leftarrow L_2 + L_3]{L_1 \leftarrow L_1} \left(\begin{array}{ccc|ccc} 1 & 0 & -2 & 1 & 0 & 0 \\ 0 & 1 & 5 & -2 & 1 & 0 \\ 0 & 0 & 10 & -4 & 1 & 1 \end{array}\right)$$

$$\xrightarrow{L_3 \leftarrow L_3/10} \left(\begin{array}{ccc|ccc} 1 & 0 & -2 & 1 & 0 & 0 \\ 0 & 1 & 5 & -2 & 1 & 0 \\ 0 & 0 & 1 & -\frac{2}{5} & \frac{1}{10} & \frac{1}{10} \end{array}\right) \xrightarrow[L_2 \leftarrow L_2 - 5L_3]{L_1 \leftarrow L_1 + 2L_3} \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ 0 & 1 & 0 & 0 & \frac{1}{2} & -\frac{1}{2} \\ 0 & 0 & 1 & -\frac{2}{5} & \frac{1}{10} & \frac{1}{10} \end{array}\right) = [\mathbb{I}_3|\mathbb{P}^{-1}]$$

   Ainsi

$$\mathbb{P}^{-1} = \begin{pmatrix} \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ 0 & \frac{1}{2} & -\frac{1}{2} \\ \frac{2}{5} & \frac{1}{10} & \frac{1}{10} \end{pmatrix}$$

4. S'il fait beau aujourd'hui, alors le vecteur d'état initial est

$$\mathbf{x}^{(0)} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

   À long terme le vecteur d'état sera $\mathbf{x} = \lim_{k \to +\infty} \mathbf{x}^{(k)}$.

   ⋆ On a déjà calculé la solution du système linéaire $(\mathbb{T} - \mathbb{I})\mathbf{x} = \mathbf{0}$, il s'agit simplement d'un vecteur propre associé à la valeur propre 1, donc si on note $\mathbf{v} = (1,2,2)^T$, on calcule $\mathbf{x} = \mathbf{v}/\sum_{i=1}^{3} v_i = (1/5, 2/5, 2/5)^T$.

2. Les systèmes étant homogènes, on n'écrira pas la matrice augmentée.

⋆ Étant donné que $\mathbb{T} = \mathbb{P}\mathbb{D}\mathbb{P}^{-1}$, alors

$$\mathbf{x} = \mathbb{P}(\lim_{k\to+\infty} \mathbb{D}^k)\mathbb{P}^{-1}\mathbf{x}^{(0)} = \begin{pmatrix} 1 & 0 & -2 \\ 2 & 1 & 1 \\ 2 & -1 & 1 \end{pmatrix} \lim_{k\to+\infty} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{4^k} & 0 \\ 0 & 0 & \frac{1}{(-4)^k} \end{pmatrix} \begin{pmatrix} \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ 0 & \frac{1}{2} & -\frac{1}{2} \\ \frac{2}{5} & \frac{1}{10} & \frac{1}{10} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & -2 \\ 2 & 1 & 1 \\ 2 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ 0 & \frac{1}{2} & -\frac{1}{2} \\ \frac{2}{5} & \frac{1}{10} & \frac{1}{10} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 0 & 0 \\ 2 & 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{5} \\ 0 \\ \frac{2}{5} \end{pmatrix} = \begin{pmatrix} \frac{1}{5} \\ \frac{2}{5} \\ \frac{2}{5} \end{pmatrix}$$

À long terme, il y a une probabilité de 20% d'avoir un beau jour, 40% d'avoir de la pluie et 40% d'avoir de la neige.

Exercice 1.92 (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 $\mathbb{A}^T\mathbb{A}$.

	Valeurs propres	Vecteurs propres unitaires
$\mathbb{A}\mathbb{A}^T = \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}$	$\lambda_1 = 9 > \lambda_2 = 1$	$\mathbb{U} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$
$\mathbb{A}^T\mathbb{A} = \begin{pmatrix} 5 & 4 & 0 \\ 4 & 5 & 0 \\ 0 & 0 & 0 \end{pmatrix}$	$\lambda_1 = 9 > \lambda_2 = 1 > \lambda_3 = 0$	$\mathbb{V} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & \sqrt{2} \end{pmatrix}$

Donc $\sigma_1^2 = 9$, $\sigma_2^2 = 1$ et

$$\mathbb{A} = \mathbb{U}\mathbb{S}\mathbb{V}^T = \underbrace{\begin{pmatrix} \mathbf{u}_1 & \dots & \mathbf{u}_r & \mathbf{u}_{r+1} & \dots & \mathbf{u}_n \end{pmatrix}}_{\in\mathbb{R}^{n\times n}} \underbrace{\begin{pmatrix} \sigma_1 & & & & & \\ & \ddots & & & & \\ & & \sigma_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix}}_{\in\mathbb{R}^{n\times p}} \underbrace{\begin{pmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_r^T \\ \mathbf{v}_{r+1}^T \\ \vdots \\ \mathbf{v}_p^T \end{pmatrix}}_{\in\mathbb{R}^{p\times p}}$$

$$= \underbrace{\begin{pmatrix} \mathbf{u}_1 & \dots & \mathbf{u}_r \end{pmatrix}}_{\in\mathbb{R}^{n\times r}} \underbrace{\begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{pmatrix}}_{\in\mathbb{R}^{r\times r}} \underbrace{\begin{pmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_r^T \end{pmatrix}}_{\in\mathbb{R}^{r\times p}} = \sum_{i=1}^{r} \sigma_i \underbrace{\mathbf{u}_i \times \mathbf{v}_i^T}_{\in\mathbb{R}^{r\times r}}$$

devient

$$\mathbb{A} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\left(\begin{array}{cc|c} 3 & 0 & 0 \\ 0 & 1 & 0 \end{array}\right)\frac{1}{\sqrt{2}}\left(\begin{array}{ccc} 1 & 1 & 0 \\ 1 & -1 & 0 \\ \hline 0 & 0 & \sqrt{2} \end{array}\right)$$

$$\overset{r=2}{=} \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \end{pmatrix}$$

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

Notons que la décomposition n'est pas unique, par exemple avec Octave on trouve

Vecteurs propres unitaires :

$$\mathbb{U} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \qquad \mathbb{V} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & \sqrt{2} \end{pmatrix}$$

ce qui donne le même résultat (heureusement!)

```
A=[1 2 0; 2 1 0]
[n,p]=size(A)
r=min(n,p)

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)

ATA=A'*A
[VecATA,ValATA]=eig(ATA)
[ee,perm] = sort(diag(abs(ValATA)),"descend");
ValATA=diag(ee)
VecATA=VecATA(:,perm)

myS=diag(sqrt(diag(ValATA)),n,p)
myU=VecAAT
myV=VecATA

[UU,SS,VV]=svd(A)

dS=diag(SS)

AA=zeros(5,4);
for i=1:length(dS)
   temp=dS(i)*UU(:,i)*VV(i,:)
   AA+=temp
end
```

CHAPITRE 2

Méthodes de résolution numériques des systèmes linéaires

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. [1] 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,
- ⋆ 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.
Bien que la plupart des méthodes de ce chapitre soient applicables aux matrices complexes, nous restreindrons notre analyse aux matrices réelles.

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 !

2.1. Méthodes directes

2.1.1. Méthode de Gauss : rappelles

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 supérieur $\mathbb{U}\mathbf{x} = \mathbf{y}$ par remontée : d'abord on trouve $x_n = \frac{y_n}{u_{nn}}$, ensuite par récurrence on déduit les inconnues $x_{n-1}, x_{n-2}, \dots x_1$ grâce à la relation $x_i = \dfrac{1}{u_{ii}}\left(y_i - \displaystyle\sum_{j=i+1}^{n} u_{ij}x_j\right)$.

Définition 2.1 (Méthode du pivot de GAUSS)
Soit $\mathbb{A} = (a_{ij})_{\substack{1\le i\le n\\ 1\le j\le p}}$ 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.

1. On entend par opération une somme, une soustraction, un produit ou une division.

EXEMPLE
Résolution du système linéaire

$$\begin{cases} x_1 + 2x_2 + 3x_3 + 4x_4 = 1, \\ 2x_1 + 3x_2 + 4x_3 + x_4 = 2, \\ 3x_1 + 4x_2 + x_3 + 2x_4 = 3, \\ 4x_1 + x_2 + 2x_3 + 3x_4 = 4. \end{cases}$$

$$[\mathbb{A}|\mathbf{b}] = \left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 2 & 3 & 4 & 1 & 2 \\ 3 & 4 & 1 & 2 & 3 \\ 4 & 1 & 2 & 3 & 4 \end{array}\right) \xrightarrow[L_4 \leftarrow L_4 - 4L_1]{\substack{L_2 \leftarrow L_2 - 2L_1 \\ L_3 \leftarrow L_3 - 3L_1}} \left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 0 & -1 & -2 & -7 & 0 \\ 0 & -2 & -8 & -10 & 0 \\ 0 & -7 & -10 & -13 & 0 \end{array}\right)$$

$$\xrightarrow[L_4 \leftarrow L_4 - 7L_2]{\substack{L_3 \leftarrow L_3 - 2L_2 \\ L_4 \leftarrow L_4 - 7L_2}} \left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 0 & -1 & -2 & -7 & 0 \\ 0 & 0 & -4 & 4 & 0 \\ 0 & 0 & 4 & 36 & 0 \end{array}\right) \xrightarrow{L_4 \leftarrow L_4 + L_3} \left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 0 & -1 & -2 & -7 & 0 \\ 0 & 0 & -4 & 4 & 0 \\ 0 & 0 & 0 & 40 & 0 \end{array}\right)$$

donc

$$\begin{cases} x_1 + 2x_2 + 3x_3 + 4x_4 = 1, \\ -x_2 - 2x_3 - 7x_4 = 0, \\ -4x_3 + 4x_4 = 0, \\ 40x_4 = 0. \end{cases} \implies x_4 = 0,\ x_3 = 0,\ x_2 = 0,\ x_1 = 1.$$

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.

2.1.2. 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 $\mathbb{A}$.) Résoudre le système $\mathbb{A}\mathbf{x} = \mathbf{b}$ équivaut à résoudre $\mathbb{LU}\mathbf{x} = \mathbf{b}$. Si on note $\mathbf{y} \overset{\text{def}}{=} \mathbb{U}\mathbf{x}$ alors on peut commencer par résoudre le système $\mathbb{L}\mathbf{y} = \mathbf{b}$ obtenant ainsi le vecteur $\mathbf{y}$, puis on résout $\mathbb{U}\mathbf{x} = \mathbf{y}$ pour obtenir $\mathbf{x}$:

résoudre $\mathbb{A}\mathbf{x} = \mathbf{b}$ revient à résoudre deux systèmes triangulaires : d'abord le système $\mathbb{L}\mathbf{y} = \mathbf{b}$, puis $\mathbb{U}\mathbf{x} = \mathbf{y}$.

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. Les deux systèmes triangulaires sont faciles à résoudre comme on va voir ci-dessous.

Proposition 2.2 (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

$$\ell_{11} y_1 = b_1$$

ce qui donne la valeur de y_1 puisque $\ell_{11} \neq 0$. En substituant cette valeur de y_1 dans les $n-1$ équations suivantes, on obtient un nouveau système dont les inconnues sont $y_2, \ldots, y_n$), pour lesquelles on peut faire de même. En procédant équation par équation, on calcule ainsi toutes les inconnues par l'algorithme dit *de descente* :

$$y_1 = \frac{b_1}{\ell_{11}}$$
$$y_i = \frac{1}{\ell_{ii}} \left(b_i - \sum_{k=1}^{i-1} \ell_{ik} y_k \right), \qquad \text{pour } i = 2, 3, \ldots, n$$

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

$$\sum_{i=1}^{n} 1 + 2 \sum_{i=1}^{n} (i-1) = n^2.$$

Proposition 2.3 (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* :

$$x_n = \frac{y_n}{u_{nn}}$$
$$x_i = \frac{1}{u_{ii}}\left(y_i - \sum_{k=i+1}^{n} u_{ik}x_k\right), \qquad \text{pour } i = n-1, n-2, \ldots, 1$$

Il nécessite également n^2 opérations.

Il reste à présent à trouver un algorithme qui permette le calcul effectif des facteurs $\mathbb{L}$ et $\mathbb{U}$ tels que $\mathbb{A} = \mathbb{L}\mathbb{U}$ avec $\mathbb{U}$ triangulaire supérieure et $\mathbb{L}$ triangulaire inférieure.

Proposition 2.4 (Factorisation $\mathbb{L}\mathbb{U}$ de DOLITTLE-v1)
Si on choisit de chercher une matrice $\mathbb{L}$ ayant la valeur 1 pour les n éléments diagonaux, 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 (on appelle cela la factorisation de DOLITTLE).

$\mathbb{L} \leftarrow \mathbb{I}_n$
for $k = 1$ à $n-1$ **do**
  for $i = k+1$ à n **do**
    $\ell_{ik} \leftarrow \dfrac{a_{ik}}{a_{kk}}$ {Attention, à chaque étape k, le terme a_{kk}, appelé pivot, doit être non nul!}
    for $j = 1$ à n **do**
      $a_{ij} \leftarrow a_{ij} - \ell_{ik}a_{kj}$ {Ligne i, colonnes $j = 1 \ldots n$: **il s'agit de u_{ik} mémorisé dans a_{ik}**}
    end for
  end for
end for
À la fin la matrice $\mathbb{A}$ est triangulaire supérieure et on pose $\mathbb{U} = \mathbb{A}$

Amélioration : à chaque 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 $\mathbb{A}$. Autrement dit, comme on sait qu'à l'étape k on a $a_{ik} = 0$ pour $i > k$, au lieu d'avoir des 0 on va y mémoriser les coefficients ℓ_{ik}.

Proposition 2.5 (Factorisation $\mathbb{L}\mathbb{U}$ de DOLITTLE-v2)
for $k = 1$ à $n-1$ **do**
  for $i = k+1$ à n **do**
    $a_{ik} \leftarrow \dfrac{a_{ik}}{a_{kk}}$ {Ligne i, colonnes $j = 1 \ldots k$: il s'agit de ℓ_{ik} mémorisé dans a_{ik}}
    for $j = k+1$ à n **do**
      $a_{ij} \leftarrow a_{ij} - a_{ik}a_{kj}$ {Ligne i, colonnes $j = k+1 \ldots n$: il s'agit de u_{ik} mémorisé dans a_{ik}}
    end for
  end for
end for
À la fin on posera :
$\mathbb{U} = \mathbb{O}_n$ puis $\mathbb{U}$ = partie triangulaire supérieure de $\mathbb{A}$
$\mathbb{L} = \mathbb{I}_n$ puis $\mathbb{L}$ = partie triangulaire strictement inférieure de $\mathbb{A}$

Proposition 2.6
Pour une matrice quelconque $\mathbb{A} \in \mathbb{R}^{n\times n}$, la factorisation $\mathbb{L}\mathbb{U}$ existe et est unique si et seulement si les sous-matrices principales $\mathbb{A}_i$ de $\mathbb{A}$ d'ordre $i = 1, \ldots, n-1$ (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 2.7
Si la matrice $\mathbb{A} \in \mathbb{R}^{n\times n}$ est symétrique et définie positive ou si est à diagonale dominante [2] alors la factorisation $\mathbb{L}\mathbb{U}$ existe et

2. $\mathbb{A} \in \mathbb{R}^{n\times n}$ est
   - ⋆ symétrique si $a_{ij} = a_{ji}$ pour tout $i, j = 1, \ldots, n$,
   - ⋆ définie positive si pour tout vecteurs $\mathbf{x} \in \mathbb{R}^n$ avec $\mathbf{x} \neq \mathbf{0}$, $\mathbf{x}^T\mathbb{A}\mathbf{x} > 0$,
   - ⋆ à diagonale dominante par lignes si $|a_{ii}| \geq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|$, pour $i = 1, \ldots, n$ (à diagonale dominante stricte par lignes si l'inégalité est stricte),
   - ⋆ à diagonale dominante par colonnes si $|a_{ii}| \geq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ji}|$, pour $i = 1, \ldots, n$ (à diagonale dominante stricte par colonnes si l'inégalité est stricte),

est unique.

Une technique qui permet d'effectuer la factorisation $\mathbb{LU}$ 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 2.8 (Algorithme de GAUSS avec pivot partiel)
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 permuter 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

$\mathbb{P} \leftarrow \mathbb{I}_n$
for $k = 1$ à $n-1$ **do**
  Dans la colonne k de $\mathbb{A}$, pour les lignes $i \geq k$ on cherche le coefficient maximal en valeur absolu et on échange sa ligne avec la ligne k, on fait la même opération sur la matrice $\mathbb{P}$
  for $i = k+1$ à n **do**
    $a_{ik} \leftarrow \dfrac{a_{ik}}{a_{kk}}$
    for $j = k+1$ à n **do**
      $a_{ij} \leftarrow a_{ij} - a_{ik}a_{kj}$
    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 $\mathbb{L}\mathbf{y} = \mathbb{P}\mathbf{b}$ puis le système triangulaire supérieure $\mathbb{U}\mathbf{x} = \mathbf{y}$.

Propriété 2.9 (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 2.10 (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

$$\mathbb{A}\mathbf{x}_i = \mathbf{e}_i, \qquad \text{pour } i = 1, \dots, n.$$

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

$$\mathbb{L}\mathbf{y}_i = \mathbb{P}\mathbf{e}_i, \qquad \mathbb{U}\mathbf{x}_i = \mathbf{y}_i, \qquad \text{pour } i = 1, \dots, n.$$

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

1. Résolution par la méthode du pivot de GAUSS du premier système

$$[\mathbb{A}|\mathbf{b}] = \left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 2 & 3 & 4 & 1 & 2 \\ 3 & 4 & 1 & 2 & 3 \\ 4 & 1 & 2 & 3 & 4 \end{array}\right) \xrightarrow[L_4 \leftarrow L_4 - 4L_1]{\substack{L_2 \leftarrow L_2 - 2L_1 \\ L_3 \leftarrow L_3 - 3L_1}} \left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 0 & -1 & -2 & -7 & 0 \\ 0 & -2 & -8 & -10 & 0 \\ 0 & -7 & -10 & -13 & 0 \end{array}\right) \xrightarrow[L_4 \leftarrow L_4 - 7L_2]{L_3 \leftarrow L_3 - 2L_2} \left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 0 & -1 & -2 & -7 & 0 \\ 0 & 0 & -4 & 4 & 0 \\ 0 & 0 & 4 & 36 & 0 \end{array}\right)$$

$$\xrightarrow{L_4 \leftarrow L_4 + L_3} \left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 0 & -1 & -2 & -7 & 0 \\ 0 & 0 & -4 & 4 & 0 \\ 0 & 0 & 0 & 40 & 0 \end{array}\right)$$

donc

$$x_4 = 0, \quad x_3 = 0, \quad x_2 = 0, \quad x_1 = 1.$$

Résolution par la méthode du pivot de GAUSS du second système

$$(\mathbb{A}|\mathbf{b}) = \left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 10 \\ 2 & 3 & 4 & 1 & 10 \\ 3 & 4 & 1 & 2 & 10 \\ 4 & 1 & 2 & 3 & 10 \end{array}\right) \xrightarrow[L_4 \leftarrow L_4 - 4L_1]{\substack{L_2 \leftarrow L_2 - 2L_1 \\ L_3 \leftarrow L_3 - 3L_1}} \left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 10 \\ 0 & -1 & -2 & -7 & -10 \\ 0 & -2 & -8 & -10 & -20 \\ 0 & -7 & -10 & -13 & -30 \end{array}\right) \xrightarrow[L_4 \leftarrow L_4 - 7L_2]{L_3 \leftarrow L_3 - 2L_2} \left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 10 \\ 0 & -1 & -2 & -7 & -10 \\ 0 & 0 & -4 & 4 & 0 \\ 0 & 0 & 4 & 36 & 40 \end{array}\right)$$

$$\xrightarrow{L_4 \leftarrow L_4 + L_3} \left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 10 \\ 0 & -1 & -2 & -7 & -10 \\ 0 & 0 & -4 & 4 & 0 \\ 0 & 0 & 0 & 40 & 40 \end{array}\right)$$

donc

$$\begin{cases} x_1 + 2x_2 + 3x_3 + 4x_4 = 10 \\ -x_2 - 2x_3 - 7x_4 = -10 \\ -4x_3 + 4x_4 = 0 \\ 40x_4 = 40 \end{cases} \quad \Longrightarrow \quad x_4 = 1, \quad x_3 = 1, \quad x_2 = 1, \quad x_1 = 1.$$

2. Factorisation de la matrice $\mathbb{A}$ (on mémorise $\mathbb{L}$ privée de sa diagonale dans la partie triangulaire inférieure de $\mathbb{A}$ car dans cette partie il n'y a que des 0 qui vont apparaître) :

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \\ 3 & 4 & 1 & 2 \\ 4 & 1 & 2 & 3 \end{pmatrix} \xrightarrow[L_4 \leftarrow L_4 - 4L_1]{\substack{L_2 \leftarrow L_2 - 2L_1 \\ L_3 \leftarrow L_3 - 3L_1}} \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & -1 & -2 & -7 \\ 3 & -2 & -8 & -10 \\ 4 & -7 & -10 & -13 \end{pmatrix} \xrightarrow[L_4 \leftarrow L_4 - 7L_2]{L_3 \leftarrow L_3 - 2L_2} \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & -1 & -2 & -7 \\ 3 & 2 & -4 & 4 \\ 4 & 7 & 4 & 36 \end{pmatrix} \xrightarrow{L_4 \leftarrow L_4 + L_3} \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & -1 & -2 & -7 \\ 3 & 2 & -4 & 4 \\ 4 & 7 & -1 & 40 \end{pmatrix}$$

donc

$$\mathbb{L} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 2 & 1 & 0 \\ 4 & 7 & -1 & 1 \end{pmatrix} \qquad \mathbb{U} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & -1 & -2 & -7 \\ 0 & 0 & -4 & 4 \\ 0 & 0 & 0 & 40 \end{pmatrix}$$

Pour résoudre le premier système linéaire on résout les systèmes triangulaires $\mathbb{L}\mathbf{y} = \mathbf{b}$

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 2 & 1 & 0 \\ 4 & 7 & -1 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} \quad \Longrightarrow \quad y_1 = 1, \quad y_2 = 0, \quad y_3 = 0, \quad y_4 = 0$$

et $\mathbb{U}\mathbf{x} = \mathbf{y}$

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & -1 & -2 & -7 \\ 0 & 0 & -4 & 4 \\ 0 & 0 & 0 & 40 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad \Longrightarrow \quad x_4 = 0, \quad x_3 = 0, \quad x_2 = 0, \quad x_1 = 1.$$

Pour résoudre le second système linéaire on résout les systèmes triangulaires $\mathbb{L}\mathbf{y} = \mathbf{b}$

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 2 & 1 & 0 \\ 4 & 7 & -1 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 10 \\ 10 \\ 10 \\ 10 \end{pmatrix} \quad \Longrightarrow \quad y_1 = 10, \quad y_2 = -10, \quad y_3 = 0, \quad y_4 = 40$$

et $\mathbb{U}\mathbf{x} = \mathbf{y}$

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & -1 & -2 & -7 \\ 0 & 0 & -4 & 4 \\ 0 & 0 & 0 & 40 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 10 \\ -10 \\ 0 \\ 40 \end{pmatrix} \qquad \Longrightarrow \qquad x_4 = 1, \quad x_3 = 1, \quad x_2 = 1, \quad x_1 = 1.$$

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

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \\ 3 & 4 & 1 & 2 \\ 4 & 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \text{ i.e. } \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 2 & 1 & 0 \\ 4 & 7 & -1 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \Longrightarrow \begin{pmatrix} 1 \\ -2 \\ 1 \\ 11 \end{pmatrix} \text{ puis } \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & -1 & -2 & -7 \\ 0 & 0 & -4 & 4 \\ 0 & 0 & 0 & 40 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ -2 \\ 1 \\ 11 \end{pmatrix} \Longrightarrow \begin{pmatrix} {}^{-9}/_{40} \\ {}^{1}/_{40} \\ {}^{1}/_{40} \\ {}^{11}/_{40} \end{pmatrix}$$

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \\ 3 & 4 & 1 & 2 \\ 4 & 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \text{ i.e. } \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 2 & 1 & 0 \\ 4 & 7 & -1 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \Longrightarrow \begin{pmatrix} 0 \\ 1 \\ -2 \\ -9 \end{pmatrix} \text{ puis } \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & -1 & -2 & -7 \\ 0 & 0 & -4 & 4 \\ 0 & 0 & 0 & 40 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ -2 \\ -9 \end{pmatrix} \Longrightarrow \begin{pmatrix} {}^{1}/_{40} \\ {}^{1}/_{40} \\ {}^{11}/_{40} \\ {}^{-9}/_{40} \end{pmatrix}$$

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \\ 3 & 4 & 1 & 2 \\ 4 & 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \text{ i.e. } \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 2 & 1 & 0 \\ 4 & 7 & -1 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \Longrightarrow \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} \text{ puis } \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & -1 & -2 & -7 \\ 0 & 0 & -4 & 4 \\ 0 & 0 & 0 & 40 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} \Longrightarrow \begin{pmatrix} {}^{1}/_{40} \\ {}^{11}/_{40} \\ {}^{-9}/_{40} \\ {}^{1}/_{40} \end{pmatrix}$$

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \\ 3 & 4 & 1 & 2 \\ 4 & 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \text{ i.e. } \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 2 & 1 & 0 \\ 4 & 7 & -1 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \Longrightarrow \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \text{ puis } \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & -1 & -2 & -7 \\ 0 & 0 & -4 & 4 \\ 0 & 0 & 0 & 40 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \Longrightarrow \begin{pmatrix} {}^{11}/_{40} \\ {}^{-9}/_{40} \\ {}^{1}/_{40} \\ {}^{1}/_{40} \end{pmatrix}$$

et finalement

$$\mathbb{A}^{-1} = \begin{pmatrix} {}^{-9}/_{40} & {}^{1}/_{40} & {}^{1}/_{40} & {}^{11}/_{40} \\ {}^{1}/_{40} & {}^{1}/_{40} & {}^{11}/_{40} & {}^{-9}/_{40} \\ {}^{1}/_{40} & {}^{11}/_{40} & {}^{-9}/_{40} & {}^{1}/_{40} \\ {}^{11}/_{40} & {}^{-9}/_{40} & {}^{1}/_{40} & {}^{1}/_{40} \end{pmatrix} = \frac{1}{40} \begin{pmatrix} -9 & 1 & 1 & 11 \\ 1 & 1 & 11 & -9 \\ 11 & 11 & -9 & 1 \\ 11 & -9 & 1 & 1 \end{pmatrix}.$$

2.2. Méthodes itératives

On n'a décrit qu'un seul algorithme de résolution, l'algorithme de GAUSS (version de DOLITTLE). 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 $\mathbf{x}^{(k)} = (x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)})^T \in \mathbb{R}^n$ convergent vers le vecteur solution exacte $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ pour tout vecteur initiale $\mathbf{x}^{(0)} = (x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)})^T \in \mathbb{R}^n$ lorsque k tend vers $+\infty$, c'est-à-dire

$$\lim_{k\to+\infty} \mathbf{x}^{(k)} = \mathbf{x}.$$

Dans ces notes on ne verra que deux méthodes itératives :

- ⋆ la méthode de JACOBI,
- ⋆ la méthode de GAUSS-SEIDEL.

Si $\mathbf{x}$ est solution du système linéaire $\mathbb{A}\mathbf{x} = \mathbf{b}$, pour toute ligne $i = 1, 2, \ldots, n$ on a

$$\underbrace{a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{i,i-1}x_{i-1}}_{\sum_{j=1}^{i-1} a_{ij}x_j} + a_{ii}x_i + \underbrace{a_{i,i+1}x_{i+1} + \cdots + a_{in}x_n}_{\sum_{j=i+1}^{n} a_{ij}x_j} = b_i$$

donc

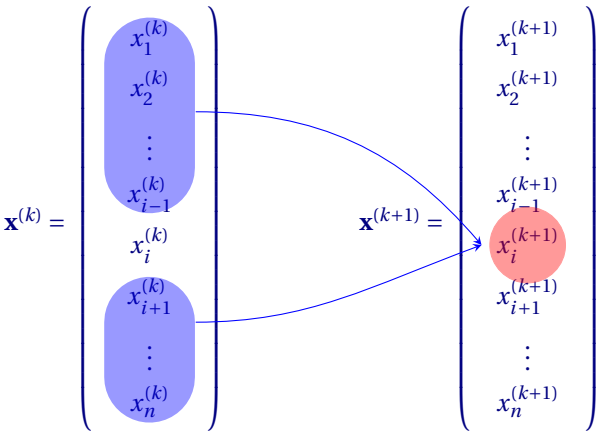
$$x_i = \frac{b_i - \sum_{j=1}^{i-1} a_{ij}x_j - \sum_{j=i+1}^{n} a_{ij}x_j}{a_{ii}} = \frac{b_i - \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}x_j}{a_{ii}}.$$

EXEMPLE

$$\begin{cases} 2x_1+x_2+x_3+4x_4=5 \\ x_1+7x_2+x_3+4x_4=9 \\ x_1+x_2+8x_3+4x_4=10 \\ 3x_1+x_2+x_3+10x_4=6 \end{cases} \Longrightarrow \begin{cases} x_1=\frac{5-\left(x_2+x_3+4x_4\right)}{2}=\frac{5-\left(x_2+x_3+4x_4\right)}{2} \\ x_2=\frac{9-\left(x_1\right)-\left(x_3+4x_4\right)}{7}=\frac{9-\left(x_1+x_3+4x_4\right)}{7} \\ x_3=\frac{10-\left(x_1+x_2\right)-\left(4x_4\right)}{8}=\frac{10-\left(x_1+x_2+4x_4\right)}{8} \\ x_4=\frac{6-\left(3x_1+x_2+x_3\right)}{10}=\frac{6-\left(3x_1+x_2+x_3\right)}{10} \end{cases}$$

Définition 2.11 (Méthode de JACOBI)
Soit $\mathbf{x}^{(0)}=(x_1^{(0)},x_2^{(0)},\ldots,x_n^{(0)})^T\in\mathbb{R}^n$ 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 :

$$x_i^{(k+1)}=\frac{b_i-\sum\limits_{\substack{j=1\\ j\neq i}}^{n}a_{ij}x_j^{(k)}}{a_{ii}},\qquad i=1,\ldots,n$$

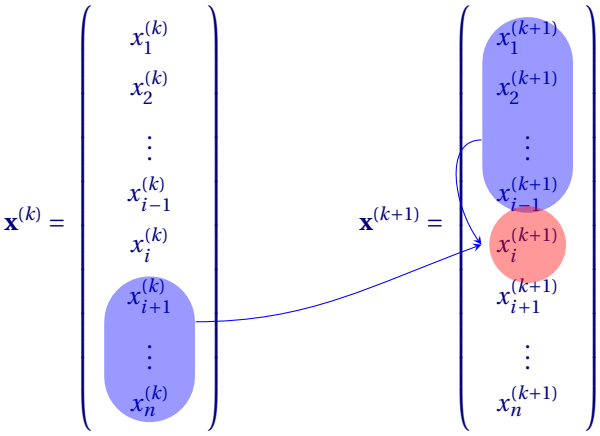


Proposition 2.12
Si la matrice $\mathbb{A}$ est à diagonale dominante stricte, la méthode de JACOBI converge.

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 2.13 (Méthode de GAUSS-SIDEL)
Soit $\mathbf{x}^{(0)}=(x_1^{(0)},x_2^{(0)},\ldots,x_n^{(0)})^T\in\mathbb{R}^n$ 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 :

$$x_i^{(k+1)}=\frac{b_i-\sum\limits_{j=1}^{i-1}a_{ij}x_j^{(k+1)}-\sum\limits_{j=i+1}^{n}a_{ij}x_j^{(k)}}{a_{ii}},\qquad i=1,\ldots,n$$



Proposition 2.14
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 2.15
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

$$\begin{cases} x = \frac{4-(2y+z)}{4} = 1 - \frac{y}{2} - \frac{z}{4}, \\ y = \frac{2-(-x+0z)}{2} = 1 + \frac{x}{2}, \\ z = \frac{9-(2x+y)}{4} = \frac{9}{4} - \frac{x}{2} - \frac{y}{4}. \end{cases}$$

Soit $\mathbf{x}^{(0)} = (0,0,0)$ le vecteur initial.

- En calculant les itérées avec la méthode de JACOBI on trouve

$$\mathbf{x}^{(1)} = \begin{pmatrix} 1 - \frac{0}{2} - \frac{0}{4} \\ 1 + \frac{0}{2} \\ \frac{9}{4} - \frac{0}{2} - \frac{0}{4} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 9/4 \end{pmatrix}, \qquad \mathbf{x}^{(2)} = \begin{pmatrix} 1 - \frac{1}{2} - \frac{9/4}{4} \\ 1 + \frac{1}{2} \\ \frac{9}{4} - \frac{1}{2} - \frac{1}{4} \end{pmatrix} = \begin{pmatrix} -1/16 \\ 3/2 \\ 3/2 \end{pmatrix},$$

$$\mathbf{x}^{(3)} = \begin{pmatrix} 1 - \frac{3/2}{2} - \frac{3/2}{4} \\ 1 + \frac{-1/16}{2} \\ \frac{9}{4} - \frac{-1/16}{2} - \frac{3/2}{4} \end{pmatrix} = \begin{pmatrix} -1/8 \\ -1/32 \\ 61/32 \end{pmatrix}, \qquad \mathbf{x}^{(4)} = \begin{pmatrix} 1 - \frac{-1/32}{2} - \frac{61/32}{4} \\ 1 + \frac{-1/8}{2} \\ \frac{9}{4} - \frac{-1/8}{2} - \frac{-1/32}{4} \end{pmatrix} = \begin{pmatrix} 5/128 \\ 15/16 \\ 265/128 \end{pmatrix}.$$

  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

$$\mathbf{x}^{(1)} = \begin{pmatrix} 1 - \frac{0}{2} - \frac{0}{4} \\ 1 + \frac{1}{2} \\ \frac{9}{4} - \frac{1}{2} - \frac{3/2}{4} \end{pmatrix} = \begin{pmatrix} 1 \\ 3/2 \\ 11/8 \end{pmatrix}, \quad \mathbf{x}^{(2)} = \begin{pmatrix} 1 - \frac{3/2}{2} - \frac{11/8}{4} \\ 1 + \frac{-3/32}{2} \\ \frac{9}{4} - \frac{-3/32}{2} - \frac{61/64}{4} \end{pmatrix} = \begin{pmatrix} -3/32 \\ 61/64 \\ 527/256 \end{pmatrix}, \quad \mathbf{x}^{(3)} = \begin{pmatrix} 1 - \frac{-3/32}{2} - \frac{61/64}{4} \\ 1 + \frac{9/1024}{2} \\ \frac{9}{4} - \frac{9/1024}{2} - \frac{2047/2048}{4} \end{pmatrix} = \begin{pmatrix} 9/1024 \\ 2047/2048 \\ 16349/8192 \end{pmatrix},$$

  La suite $\mathbf{x}^{(k)}$ converge vers $(0,1,2)$ la solution du système.

Remarque (Quand doit-on arrêter une méthode itérative?)
En théorie, il faudrait effectuer un nombre infini d'itérations pour obtenir la solution exacte d'un système linéaire avec une méthode itérative. En pratique, ce n'est ni nécessaire, ni raisonnable (même si effectivement le nombre d'itérations pour obtenir la solution avec la précision machine peut être très élevé pour de grands systèmes). En effet, ce n'est en général pas d'une solution exacte dont on a besoin, mais plutôt d'une valeur $\mathbf{x}^{(k)}$ qui approche la solution exacte avec une erreur inférieure à une tolérance toll fixée. Mais comme l'erreur est elle-même inconnue (puisqu'elle dépend de la solution exacte), on a besoin d'un estimateur d'erreur a posteriori qui donne une estimation de l'erreur à partir de quantités calculées au cours de la résolution.
Un premier estimateur est donné par le résidu :

$$\mathbf{r}^{(k)} \stackrel{\text{def}}{=} \mathbb{A}\mathbf{x}^{(k)} - \mathbf{b}.$$

Ainsi, on peut décider de stopper les itérations à la première étape $k_{\min}$ pour laquelle

$$\|\mathbf{r}^{(k)}\| \leq \text{toll}\|\mathbf{b}\|.$$

Le contrôle par le résidu n'est pertinent que pour les matrices dont le conditionnement n'est pas trop grand.
Un autre estimateur est donné par l'incrément

$$\delta(\mathbf{x}) \stackrel{\text{def}}{=} \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}.$$

On peut choisir de stopper les itérations à la première étape $k_{\min}$ pour laquelle

$$\|\delta(\mathbf{x})\| \leq \text{toll}.$$

2.3. Quelle est la précision de la solution d'un système linéaire ?

Le produit $\mathbb{L}\mathbb{U}$ n'est pas exactement égal à $\mathbb{A}$ en pratique, à cause des erreurs d'arrondi. Bien que la stratégie du pivot atténue ces erreurs, le résultat n'est pas toujours très satisfaisant.

EXEMPLE
Le couple $x_1 = x_2 = 1$ est solution du système de deux équations à deux inconnues suivant :

$$\begin{cases} 3.218613x_1 + 6.327917x_2 = 10.546530, \\ 3.141592x_1 + 4.712390x_2 = 7.853982. \end{cases}$$

Considérons maintenant un système d'équations "voisin" (le carré indique un changement de décimale) :

$$\begin{cases} 3.21861\boxed{1}x_1 + 6.327917x_2 = 10.546530, \\ 3.14159\boxed{4}x_1 + 4.712390x_2 = 7.85398\boxed{0}. \end{cases}$$

Sa solution est donnée par $x_1 = -5$, $x_2 = 5$.
On voit que, bien que ces deux systèmes soient voisins, leurs solutions sont très différentes : on parle dans ce cas de *systèmes mal conditionnés*.
Géométriquement, ces deux systèmes peuvent être vus comme la recherche du point d'intersection entre deux droites. Si une petite perturbation sur la pente donne une grande différence sur la solution, alors les deux droites sont "presque" parallèles.

Résoudre un système mal conditionné avec un ordinateur peut être une affaire délicate si l'ordinateur calcule avec trop peu de chiffres significatifs. Dans l'exemple précédent nous observons que, si l'ordinateur ne retient que 6 chiffres significatifs, il est complètement inespéré d'obtenir une solution raisonnablement proche de la solution !

Considérons un système linéaire (non singulier) $\mathbb{A}\mathbf{x} = \mathbf{b}$ et le système linéaire perturbé $\mathbb{A}\mathbf{y} = \mathbf{b} + \delta\mathbf{b}$ où $\delta\mathbf{b}$ est une petite perturbation de $\mathbf{b}$. Par linéarité la solution $\mathbf{y}$ du système perturbé est liée à la solution $\mathbf{x}$ du système non perturbé par la relation $\mathbf{y} = \mathbf{x} + \delta\mathbf{x}$ 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 $\mathbb{A}$.

Définition 2.16 (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 normes matricielles suivantes ($p \geq 1$) :

$$\|\mathbb{A}\|_1 = \max_{j=1,\dots,n} \sum_{i=1}^{n} |a_{ij}|, \qquad \|\mathbb{A}\|_2 = \sqrt{\lambda_{\max}(\mathbb{A}^T\mathbb{A})}, \qquad \|\mathbb{A}\|_\infty = \max_{i=1,\dots,n} \sum_{j=1}^{n} |a_{ij}|.$$

EXEMPLE

$$\mathbb{A} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} \quad \Longrightarrow \quad \begin{array}{l} \|\mathbb{A}\|_1 = \max(1+4+7, 2+5+8, 3+6+9) = \max(12, 15, 18) = 18, \\ \|\mathbb{A}\|_\infty = \max(1+2+3, 4+5+6, 7+8+9) = \max(6, 15, 24) = 24. \\ \|\mathbb{A}\|_2 = \sqrt{\varrho(\mathbb{A}\mathbb{A}^T)} = 16.848 \end{array}$$

```
A=[1 2 3;4 5 6; 7,8 9]
% norm_1
max(sum(abs(A)))
norm(A,1)
% norm_inf
max(sum(abs(A')))
norm(A,inf)
% norm_2
sqrt(max(abs(eig(A'*A))))
norm(A,2)
% conditionnement en norme 2
cond(A)
```

Si $K(\mathbb{A})$ est "petit", c'est-à-dire de l'ordre de l'unité, on dit que $\mathbb{A}$ est bien conditionnée. Dans ce cas, des erreurs sur les données induisent des erreurs du même ordre de grandeur sur la solution. Cette propriété intéressante n'est plus vérifiée par les matrices mal conditionnées.
Si $\|\delta\mathbf{b}\|/\|\mathbf{b}\|$ est de l'ordre de la précision relative $\eta = 10^{-p}$ du calculateur, alors l'erreur relative sur la solution $\|\delta\mathbf{x}\|/\|\mathbf{x}\|$ 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 calcul la solution du système linéaire avec un ordinateur à p chiffres significatifs en valeur décimale, on ne pourra pas garantir à priori plus de

$$E(p - \log_{10}(K(\mathbb{A})))$$

chiffres significatifs sur la solution.

Nota bene : 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.

Remarque (Cas particulier)
Si $\mathbb{A}$ est symétrique et définie positive [3],

$$K_2(\mathbb{A}) = \|\mathbb{A}\|_2\|\mathbb{A}^{-1}\|_2 = \frac{\lambda_{\max}}{\lambda_{\min}}$$

où $\lambda_{\max}$ (resp. $\lambda_{\min}$) est la plus grande (resp. petite) valeur propre de $\mathbb{A}$.

Une étude analogue peut être réalisée pour des petites perturbations des coefficients de la matrice $\mathbb{A}$.

2.4. Exercices

Factorisation $\mathbb{LU}$ et systèmes linéaires carrés

Exercice 2.1
Soit $\mathbb{A}$ une matrice, $\mathbb{A} \in \mathscr{M}_{n,n}(\mathbb{R})$.

1. Rappeler les conditions nécessaires et suffisantes pour l'existence d'une factorisation $\mathbb{LU}$ de la matrice $\mathbb{A}$ et préciser les définitions de $\mathbb{L}$ et $\mathbb{U}$.
2. On suppose $\mathbb{L}$ et $\mathbb{U}$ construites (*i.e.* on dispose de tous les coefficients $\ell_{i,j}$ et $u_{i,j}$ de $\mathbb{L}$ et $\mathbb{U}$), écrire l'algorithme de

3. $\mathbb{A} \in \mathbb{R}^{n\times n}$ est
   - ⋆ symétrique si $a_{ij} = a_{ji}$ pour tout $i, j = 1, \ldots, n$,
   - ⋆ définie positive si pour tout vecteurs $\mathbf{x} \in \mathbb{R}^n$ avec $\mathbf{x} \neq \mathbf{0}$, $\mathbf{x}^T\mathbb{A}\mathbf{x} > 0$.

résolution de $\mathbb{A}\mathbf{x} = \mathbf{b}$, avec $\mathbf{b} \in \mathscr{M}_{n,1}(\mathbb{R})$ 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 $\mathbb{L}$ et $\mathbb{U}$ de la factorisation $\mathbb{LU}$.

Correction

1. Pour une matrice quelconque $\mathbb{A} \in \mathscr{M}_{n,n}(\mathbb{R})$, la factorisation $\mathbb{LU}$ (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). 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 $\mathbb{LU}$ 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.
2. Une fois calculées les matrices $\mathbb{L}$ et $\mathbb{U}$, résoudre le système linéaire $\mathbb{A}\mathbf{x} = \mathbf{b}$, avec $\mathbf{b} \in \mathscr{M}_{n,1}(\mathbb{R})$ donné consiste simplement à résoudre successivement
   2.1. le système triangulaire inférieur $\mathbb{L}\mathbf{y} = \mathbf{b}$ par l'algorithme
$$y_1 = b_1, \qquad y_i = b_i - \sum_{j=1}^{i-1} \ell_{ij} y_j, \qquad i = 2,\dots,n$$
   Rappel : $\ell_{ii} = 1$ dans la factorisation de Dolittle.
   2.2. le système triangulaire supérieure $\mathbb{U}\mathbf{x} = \mathbf{y}$ par l'algorithme
$$x_n = \frac{y_n}{u_{nn}}, \qquad x_i = \frac{1}{u_{ii}}\left(y_i - \sum_{j=i+1}^{n} u_{ij} x_j\right), \qquad j = n-1,\dots,1$$
3. Factorisation :
$$\begin{pmatrix} 3 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{pmatrix} \xrightarrow[L_3 \leftarrow L_3 - \frac{-1}{3} L_1]{L_2 \leftarrow L_2 - \frac{-1}{3} L_1} \begin{pmatrix} 3 & -1 & -1 \\ 0 & 8/3 & -4/3 \\ 0 & -4/3 & 8/3 \end{pmatrix} \xrightarrow{L_3 \leftarrow L_3 - \frac{-4/3}{8/3} L_2} \begin{pmatrix} 3 & -1 & -1 \\ 0 & 8/3 & -4/3 \\ 0 & 0 & 2 \end{pmatrix}.$$
Par conséquent
$$\mathbb{L} = \begin{pmatrix} 1 & 0 & 0 \\ -1/3 & 1 & 0 \\ -1/3 & -1/2 & 1 \end{pmatrix} \quad \text{et} \quad \mathbb{U} = \begin{pmatrix} 3 & -1 & -1 \\ 0 & 8/3 & -4/3 \\ 0 & 0 & 2 \end{pmatrix}.$$

Exercice 2.2

Résoudre les systèmes linéaires suivants :
$$\begin{cases} x - 1 - 5x_2 - 7x_3 = 3 \\ 2x_1 - 13x_2 - 18x_3 = 3 \\ 3x_1 - 27x_2 - 36x_3 = 3 \end{cases} \quad \text{et} \quad \begin{cases} x - 1 - 5x_2 - 7x_3 = 6 \\ 2x_1 - 13x_2 - 18x_3 = 0 \\ 3x_1 - 27x_2 - 36x_3 = -3 \end{cases} \quad \text{et} \quad \begin{cases} x - 1 - 5x_2 - 7x_3 = 0 \\ 2x_1 - 13x_2 - 18x_3 = 3 \\ 3x_1 - 27x_2 - 36x_3 = 6. \end{cases}$$

Correction

Le trois systèmes s'écrivent sous forme matricielle
$$\begin{pmatrix} 1 & -5 & -7 \\ 2 & -13 & -18 \\ 3 & -27 & -36 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \\ 3 \end{pmatrix} \quad \text{et} \quad \begin{pmatrix} 1 & -5 & -7 \\ 2 & -13 & -18 \\ 3 & -27 & -36 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 6 \\ 0 \\ -3 \end{pmatrix} \quad \text{et} \quad \begin{pmatrix} 1 & -5 & -7 \\ 2 & -13 & -18 \\ 3 & -27 & -36 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 3 \\ 6 \end{pmatrix}$$
On remarque que seul le terme source change. On calcul d'abord la décomposition $\mathbb{LU}$ de la matrice $\mathbb{A}$:
$$\begin{pmatrix} 1 & -5 & -7 \\ 2 & -13 & -18 \\ 3 & -27 & -36 \end{pmatrix} \xrightarrow[L_3 \leftarrow L_3 - 3L_1]{L_2 \leftarrow L_2 - 2L_1} \begin{pmatrix} 1 & -5 & -7 \\ 0 & -3 & -4 \\ 0 & -12 & -15 \end{pmatrix} \xrightarrow{L_3 \leftarrow L_3 - 4L_2} \begin{pmatrix} 1 & -5 & -7 \\ 0 & -3 & -4 \\ 0 & 0 & 1 \end{pmatrix}$$

donc

$$\mathbb{L} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 4 & 1 \end{pmatrix} \qquad \mathbb{U} = \begin{pmatrix} 1 & -5 & -7 \\ 0 & -3 & -4 \\ 0 & 0 & 1 \end{pmatrix}$$

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

$$\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 4 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \\ 3 \end{pmatrix} \implies y_1 = 3, \quad y_2 = -3, \quad y_3 = 6;$$

$$\begin{pmatrix} 1 & -5 & -7 \\ 0 & -3 & -4 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 3 \\ -3 \\ 6 \end{pmatrix} \implies x_3 = 6, \quad x_2 = -7, \quad x_1 = 10.$$

2. Pour le seconde système on a

$$\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 4 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 6 \\ 0 \\ -3 \end{pmatrix} \implies y_1 = 6, \quad y_2 = -12, \quad y_3 = 27;$$

$$\begin{pmatrix} 1 & -5 & -7 \\ 0 & -3 & -4 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 6 \\ -12 \\ 27 \end{pmatrix} \implies x_3 = 27, \quad x_2 = -32, \quad x_1 = 35.$$

3. Pour le dernier système on a

$$\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 4 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 3 \\ 6 \end{pmatrix} \implies y_1 = 0, \quad y_2 = 3, \quad y_3 = -6;$$

$$\begin{pmatrix} 1 & -5 & -7 \\ 0 & -3 & -4 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 3 \\ -6 \end{pmatrix} \implies x_3 = -6, \quad x_2 = 7, \quad x_1 = -7.$$

☆ Exercice 2.3

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 $\mathbf{b}$ est un vecteur colonne de $\mathbb{R}^n$ et $\mathbb{L}$ une matrice de $\mathbb{R}^{n\times n}$ triangulaire inférieure. On doit obtenir un vecteur colonne de $\mathbb{R}^n$ solution du système linaire $\mathbb{L}\mathbf{y} = \mathbf{b}$. Écrire un script appelé `TESTdescente.m` pour tester cette fonction sur l'exemple suivant : pour

$$\mathbb{L} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 3 & 0 \\ 4 & 5 & 6 \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} 1 \\ 8 \\ 32 \end{pmatrix}$$

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 `function x=remontee(U,y)` où $\mathbf{y}$ est un vecteur colonne de $\mathbb{R}^n$ et $\mathbb{U}$ une matrice de $\mathbb{R}^{n\times n}$ triangulaire supérieure. On doit obtenir un vecteur colonne de $\mathbb{R}^n$ solution du système

linaire $\mathbb{U}\mathbf{x}=\mathbf{y}$. Écrire un script appelé `TESTremontee.m` pour tester cette fonction sur l'exemple suivant : pour

$$\mathbb{U}=\begin{pmatrix}1&2&3\\0&4&5\\0&0&6\end{pmatrix} \qquad \mathbf{y}=\begin{pmatrix}14\\23\\18\end{pmatrix}$$

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 $\mathbb{LU}$ d'une matrice $\mathbb{A}$ par la méthode de GAUSS. La syntaxe doit être `function [L,U]=mylu(A)` où $\mathbb{A}$ est une matrice de $\mathbb{R}^{n\times n}$ non singulière et $\mathbf{b}$ est un vecteur colonne de $\mathbb{R}^n$. On doit obtenir $\mathbb{L}$ et $\mathbb{U}$ deux matrices de $\mathbb{R}^{n\times n}$ triangulaires inférieur et supérieur respectivement telles que $\mathbb{LU}=\mathbb{A}$. Écrire un script appelé `TESTmylu.m` pour tester cette fonction sur l'exemple suivant : pour

$$\mathbb{A}=\begin{pmatrix}1&0&3\\2&2&2\\3&6&4\end{pmatrix}$$

on doit obtenir

$$\mathbb{L}=\begin{pmatrix}1&0&0\\2&1&0\\3&3&1\end{pmatrix} \qquad \mathbb{U}=\begin{pmatrix}1&0&3\\0&2&-4\\0&0&7\end{pmatrix}$$

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 $\mathbb{A}\mathbf{x}=\mathbf{b}$ en utilisant la factorisation $\mathbb{LU}$ de la matrice $\mathbb{A}$ puis la résolution des systèmes linéaires $\mathbb{L}\mathbf{y}=\mathbf{b}$ et $\mathbb{U}\mathbf{x}=\mathbf{y}$. La syntaxe doit être `function [x]=syslin(A,b)` où $\mathbb{A}$ est une matrice de $\mathbb{R}^{n\times n}$ non singulière. On doit obtenir $\mathbf{x}$ un vecteur colonne de $\mathbb{R}^n$ solution du système linaire $\mathbb{A}\mathbf{x}=\mathbf{b}$. Écrire un script appelé `TESTsyslin.m` pour tester cette fonction sur l'exemple suivant : pour

$$\mathbb{A}=\begin{pmatrix}1&0&3\\2&2&2\\3&6&4\end{pmatrix} \qquad \mathbf{b}=\begin{pmatrix}4\\6\\13\end{pmatrix}$$

on doit obtenir

$$\mathbf{x}=\begin{pmatrix}1\\1\\1\end{pmatrix}$$

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 $\mathbb{A}$ en utilisant la factorisation $\mathbb{LU}$ de la matrice $\mathbb{A}$. La syntaxe doit être `function [d]=mydet(A)` où $\mathbb{A}$ est une matrice de $\mathbb{R}^{n\times n}$ non singulière. On doit obtenir $d=\det(\mathbb{A})$. 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 $\mathbb{A}^{-1}$ d'une matrice $\mathbb{A}$ en utilisant la factorisation $\mathbb{LU}$ de la matrice $\mathbb{A}$ et la résolution des $2n$ systèmes linéaires $\mathbb{L}\mathbf{y} = \mathbf{e}_j$ et $\mathbb{U}\mathbf{x} = \mathbf{y}$ avec $\mathbf{e}_j$ le vecteur $(\mathbf{e}_j)_i = \delta_{ij}$. La syntaxe doit être `function [invA]=myinv(A)` où $\mathbb{A}$ est une matrice de $\mathbb{R}^{n\times n}$ non singulière. On doit obtenir $\mathbb{A}^{-1}$ une matrice de $\mathbb{R}^{n\times n}$ telle que $\mathbb{A}^{-1}\mathbb{A} = \mathbb{A}\mathbb{A}^{-1} = \mathbb{I}_n$. Écrire un script appelé `TESTmyinv.m` pour tester cette fonction sur l'exemple suivant : pour

$$\mathbb{A} = \begin{pmatrix} 1 & 0 & 3 \\ 2 & 2 & 2 \\ 3 & 6 & 4 \end{pmatrix}$$

on doit obtenir

$$\mathbb{A}^{-1} = \begin{pmatrix} -1 & 0 & 1 \\ -1 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & -\frac{1}{2} \end{pmatrix}$$

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 $\mathbb{LU}$ d'une matrice $\mathbb{A}$ par la méthode de GAUSS avec pivot, *i.e.* $\mathbb{PA} = \mathbb{LU}$. La syntaxe doit être `function [L,U,P]=mylupivot(A)` où $\mathbb{A}$ est une matrice de $\mathbb{R}^{n\times n}$ non singulière.
   Expliquer pourquoi on ne peut pas effectuer la factorisation $\mathbb{LU}$ de la matrice

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

   mais on peut effectuer la factorisation avec pivot. Calculer cette factorisation.

8. Implémenter une fonction appelée `syslinpivot.m` permettant de résoudre le système linéaire $\mathbb{A}\mathbf{x} = \mathbf{b}$ en utilisant la factorisation $\mathbb{LU}$ avec pivot de la matrice $\mathbb{A}$ puis la résolution des systèmes linéaires $\mathbb{L}\mathbf{y} = \mathbb{P}\mathbf{b}$ et $\mathbb{U}\mathbf{x} = \mathbf{y}$. La syntaxe doit être `function [x]=syslinpivot(A,b)` où $\mathbb{A}$ est une matrice de $\mathbb{R}^{n\times n}$ non singulière. On doit obtenir $\mathbf{x}$ un vecteur colonne de $\mathbb{R}^n$ solution du système linaire $\mathbb{A}\mathbf{x} = \mathbf{b}$. Écrire un script appelé `TESTsyslinpivot.m` pour tester cette fonction sur l'exemple suivant : pour

$$\mathbb{A} = \begin{pmatrix} 1 & 1 & 3 \\ 2 & 2 & 2 \\ 3 & 6 & 4 \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} 5 \\ 6 \\ 13 \end{pmatrix}$$

   on doit obtenir

$$\mathbf{x} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

   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. Il s'agit d'implémenter la méthode de descente pour résoudre un système triangulaire inférieure $\mathbb{L}\mathbf{y} = \mathbf{b}$:

$$y_1 = \frac{b_1}{\ell_{11}}$$
$$y_i = \frac{1}{\ell_{ii}}\left(b_i - \sum_{k=1}^{i-1} \ell_{ik} y_k\right), \qquad \text{pour } i = 2,3,\dots,n$$

   Si la matrice $\mathbb{L}$ est obtenue par la méthode de factorisation de Doolittle, alors $\ell_{ii} = 1$ pour tout $i = 1,\dots,n$.

   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)-sum(L(i,1:i-1).*y(1:i-1)))/L(i,i);
  end
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 verifier notre resultat on peut
% soit comparer le resultat avec celui d'Octave
yOctave=L\b
% soit verifier que Ly=b
Ly=L*y
```

2. Il s'agit d'implémenter la méthode de remontée pour résoudre un système triangulaire supérieure $\mathbb{U}\mathbf{x} = \mathbf{y}$:

$$x_n = \frac{y_n}{u_{nn}}$$
$$x_i = \frac{1}{u_{ii}}\left(y_i - \sum_{k=i+1}^{n} u_{ik}x_k\right), \qquad \text{pour } i = n-1, n-2, \ldots, 1$$

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)-sum(U(i,i+1:n).*x(i+1:n)))/U(i,i);
  end
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 verifier notre resultat on peut
% soit comparer le resultat avec celui d'Octave
xOctave=U\y
% soit verifier que Ux=y
Ux=U*x
```

3. Il s'agit d'implémenter la méthode de factorisation qui calcule deux matrices $\mathbb{L}$ et $\mathbb{U}$ telles que $\mathbb{A} = \mathbb{L}\mathbb{U}$ avec $\mathbb{U}$ triangulaire supérieure obtenue par ma méthode de GAUSS et $\mathbb{L}$ triangulaire inférieure avec que des 1 sur la diagonale.
Version basique :

$\mathbb{L} \leftarrow \mathbb{I}_n$
for $k = 1$ à $n-1$ **do**
  for $i = k+1$ à n **do**
    $\ell_{ik} \leftarrow \dfrac{a_{ik}}{a_{kk}}$ {Attention, à chaque étape k, le terme a_{kk}, appelé pivot, doit être non nul !}
    for $j = 1$ à n **do**
      $a_{ij} \leftarrow a_{ij} - \ell_{ik}a_{kj}$ {On modifie tous les éléments de la ligne i}
    end for
  end for
end for
$\mathbb{U} = \mathbb{A}$

Dans le fichier `myluBIS.m` on écrit

```
function [L,U]=myluBIS(A)
% Factorisation de Doolittle, i.e. L(i,i)=1
   [n,m]=size(A);
   L=eye(n);
   for k=1:n-1
      for i=k+1:n
         L(i,k)=A(i,k)/A(k,k);
         A(i,:)=A(i,:)-L(i,k)*A(k,:);
      end
   end
   U=A;
end
```

Version améliorée : à chaque étape k on peut mémoriser ℓ_{ik} en a_{ik} :

for $k = 1$ à $n-1$ **do**
  for $i = k+1$ à n **do**
    $a_{ik} \leftarrow \dfrac{a_{ik}}{a_{kk}}$ {Il s'agit de ℓ_{ik} mémorisé dans a_{ik}}
    for $j = k+1$ à n **do**
      $a_{ij} \leftarrow a_{ij} - a_{ik}a_{kj}$ {On modifie tous les éléments de la ligne i}
    end for
  end for
end for

$\mathbb{U} = \mathbb{O}_n$
$\mathbb{U}$ = partie triangulaire supérieure de$\mathbb{A}$
$\mathbb{L} = \mathbb{I}_n$
$\mathbb{L}$ = partie triangulaire strictement inférieure de $\mathbb{A}$

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 ces fonctions par exemple comme suit

```
printf(...
"============================================================\n\
Test 1\n\
============================================================\n");
A=[1 0 3; 2 2 2; 3 6 4]
[L,U]=mylu(A)
[L,U]=myluBIS(A)
% Verifions notre resutat 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)
[L,U]=myluBIS(A)
LU=L*U

printf(...
"============================================================\n\
Test 3\n\
============================================================\n");
A=[1 1 3; 2 2 2; 3 6 4]
[L,U]=mylu(A)
[L,U]=myluBIS(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 qua 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
% le comparer au resultat d'Octave
dOctave=det(A)
```

6. La j-ème colonne de $\mathbb{A}^{-1}$ est solution du système linéaire $\mathbb{A}\mathbf{x} = \mathbf{e}_j$ ou $\mathbf{e}_j$ est le vecteur qui contient 1 sur la j-ème ligne est 0 ailleurs. Comme la matrice $\mathbb{A}$ est la même pour tous les systèmes linéaires, on n'utilisera pas la fonction `syslin` car cela voudrait dire factoriser n fois la matrice $\mathbb{A}$.

Dans le fichier `myinv.m` on écrit

```
function invA=myinv(A)
   [n,m]=size(A);
   [L,U]=mylu(A);
   for j=1:n
      b=zeros(n);
      b(j)=1;
      y=descente(L,b)';
      invA(:,j)=remontee(U,y)';
   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. **for** $k = 1$ à $n-1$ **do**
   Dans la colonne k, pour les lignes $i \geq k$ on cherche le coefficient maximal en valeur absolu et on échange sa ligne avec la ligne k
   for $i = k+1$ à n **do**
   $a_{ik} \leftarrow \dfrac{a_{ik}}{a_{kk}}$ {Il s'agit de ℓ_{ik} mémorisé dans a_{ik}}
   for $j = k+1$ à n **do**
   $a_{ij} \leftarrow a_{ij} - a_{ik}a_{kj}$ {On modifie tous les éléments de la ligne i}
   end for
   end for
   end for
   $\mathbb{U} = \mathbb{O}_n$
   $\mathbb{U}$ = partie triangulaire supérieure de$\mathbb{A}$
   $\mathbb{L} = \mathbb{I}_n$
   $\mathbb{L}$ = partie triangulaire strictement inférieure de$\mathbb{A}$
   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(...
"============================================================\n\
Test 1\n\
============================================================\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(...
"============================================================\n\
Test 2\n\
============================================================\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 ecrire 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 $\mathbb{LU}$ sans pivot car $\det(\mathbb{A}_1) = 1$ mais $\det(\mathbb{A}_2) = \left(\begin{smallmatrix} 1 & 1 \\ 2 & 2 \end{smallmatrix}\right) = 0$. Cependant on peut calculer la factorisation $\mathbb{LU}$ avec pivot car la matrice n'est pas singulière.

8. Le système linéaire $\mathbb{A}\mathbf{x} = \mathbf{b}$ est équivalent au système linéaire $\mathbb{LU}\mathbf{x} = \mathbb{P}\mathbf{b}$ avec $\mathbb{L}$ et $\mathbb{U}$ telles que $\mathbb{PA} = \mathbb{LU}$ avec $\mathbb{U}$ triangulaire supérieure et $\mathbb{L}$ triangulaire inférieure obtenues par la méthode de factorisation non pas de la matrice $\mathbb{A}$ mais de la matrice $\mathbb{PA}$. On calcule alors d'abord $\mathbf{y}$ solution du système linéaire triangulaire inférieur $\mathbb{L}\mathbf{y} = \mathbb{P}\mathbf{b}$ par l'algorithme de descente, puis on calcule $\mathbf{x}$ solution du système linéaire triangulaire supérieur $\mathbb{U}\mathbf{x} = \mathbf{y}$ par l'algorithme de remontée.

Dans le fichier `syslinpivot.m` on écrit

```
function x=syslinpivot(A,b)
   [L,U,P]=mylupivot(A);
   y=descente(L,P*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=syslinpivot(A,b)
% Comparons le resultat a celui d'Octave
xOctave=A\b
% Verifions 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 resultat a celui d'Octave
xOctave=A\b
% Verifions que Ax=b
Ax=A*x
```

☆ Exercice 2.4 (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 $\mathbb{LU}$ 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 $\mathbb{LU}$ de $\mathbb{A}$ sont bidiagonales, écrivons-les sous la forme

$$\mathbb{L} = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & 0 \\ \beta_2 & 1 & \ddots & & & \vdots \\ 0 & \beta_3 & 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \beta_{n-1} & 1 & 0 \\ 0 & \dots & \dots & 0 & \beta_n & 1 \end{pmatrix}, \qquad \mathbb{U} = \begin{pmatrix} \alpha_1 & \gamma_1 & 0 & \dots & \dots & 0 \\ 0 & \alpha_2 & \gamma_2 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ \vdots & & & \ddots & \alpha_{n-1} & \gamma_{n-1} \\ 0 & \dots & \dots & \dots & 0 & \alpha_n \end{pmatrix}.$$

   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.

   Hérédité : 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 $i \le k$ 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 apparaitre 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 :

$$\mathbb{A}^{(0)} = \begin{pmatrix} a_1 & c_1 & 0 & \dots & \dots & 0 \\ b_2 & a_2 & c_2 & \ddots & & \vdots \\ 0 & b_3 & a_3 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & b_{n-1} & a_{n-1} & c_{n-1} \\ 0 & \dots & \dots & 0 & b_n & a_n \end{pmatrix}$$

$$\xrightarrow[\beta_2 = \frac{b_2}{a_1}]{L_2 \leftarrow L_2 - \beta_2 L_1} \mathbb{A}^{(1)} = \begin{pmatrix} \alpha_1 = a_1 & \gamma_1 = c_1 & 0 & \dots & \dots & 0 \\ 0 & \alpha_2 = a_2 - \beta_2 c_1 & \gamma_2 = c_2 & \ddots & & \vdots \\ 0 & b_3 & a_3 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & b_{n-1} & a_{n-1} & c_{n-1} \\ 0 & \dots & \dots & 0 & b_n & a_n \end{pmatrix}$$

$$\xrightarrow[\beta_3 = \frac{b_3}{\alpha_2}]{L_3 \leftarrow L_3 - \beta_3 L_2} \mathbb{A}^{(2)} = \begin{pmatrix} \alpha_1 = a_1 & \gamma_1 = c_1 & 0 & \dots & \dots & 0 \\ 0 & \alpha_2 = a_2 - \beta_2 c_1 & \gamma_2 = c_2 & \ddots & & \vdots \\ 0 & 0 & \alpha_3 = a_3 - \beta_3 c_2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & b_{n-1} & a_{n-1} & c_{n-1} \\ 0 & \dots & \dots & 0 & b_n & a_n \end{pmatrix} \xrightarrow[\beta_4 = \frac{b_4}{\alpha_3}]{L_4 \leftarrow L_4 - \beta_4 L_3} [\cdots]$$

$$[\cdots] \xrightarrow[\beta_n = \frac{b_n}{\alpha_n}]{L_n \leftarrow L_n - \beta_n L_{n-1}} \mathbb{A}^{(n-1)} = \begin{pmatrix} \alpha_1 = a_1 & \gamma_1 = c_1 & 0 & \dots & \dots & 0 \\ 0 & \alpha_2 = a_2 - \beta_2 c_1 & \gamma_2 = c_2 & \ddots & & \vdots \\ 0 & 0 & \alpha_3 = a_3 - \beta_3 c_2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & 0 & \alpha_{n-1} = a_{n-1} - \beta_{n-1} c_{n-2} & \gamma_{n-1} = c_{n-1} \\ 0 & \dots & \dots & 0 & 0 & \alpha_n = a_n - \beta_n c_{n-1} \end{pmatrix}$$

Donc $\gamma_i = c_i$ pour $i = 1, \dots, n$, $\alpha_1 = a_1$ et on définie par récurrence

$$\begin{cases} \beta_i &= \frac{b_i}{\alpha_{i-1}} \\ \alpha_i &= a_i - \beta_i c_{i-1} \end{cases} \qquad \text{pour } i = 2, \dots, n.$$

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 :

$$\begin{cases} y_1 = f_1, \\ y_i = f_i - \beta_i y_{i-1}, \quad \text{pour } i = 2,\dots,n, \end{cases}$$

$$\begin{cases} x_n = \frac{y_n}{\alpha_n}, \\ x_i = \frac{y_i - \gamma_i x_{i+1}}{\alpha_i}, \quad \text{pour } i = n-1,\dots,1, \end{cases}$$

i.e.

$$\begin{cases} y_1 = f_1, \\ y_i = f_i - \frac{b_i}{a_i - \beta_i c_{i-1}} y_{i-1}, \quad \text{pour } i = 2,\dots,n; \end{cases}$$

i.e.

$$\begin{cases} x_n = \frac{y_n}{\alpha_n}, \\ x_i = \frac{y_i - c_i x_{i+1}}{a_i - \beta_i c_{i-1}}, \quad \text{pour } i = n-1,\dots,2, \\ x_1 = \frac{y_i - c_i x_2}{a_1}. \end{cases}$$

Dans le fichier `syslinThomas.m` on écrit

```
function x=syslinThomas(a,b,c,f)
   [alpha,beta]=myluThomas(a,b,c);
   y=descenteThomas(beta,f)';
   x=remonteeThomas(alpha,c,y)';
end
```

et on teste cette fonction par exemple comme suit

```
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 tridigonale
% 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 2.5

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 $\mathbb{LU}$ de $\mathbb{A}$.

Correction

Factorisation $\mathbb{LU}$ de la matrice $\mathbb{A}$:

$$\begin{pmatrix} 1 & -1 & \dots & \dots & \dots & -1 \\ 0 & 1 & \ddots & & & \vdots \\ \vdots & \ddots & 1 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & 1 & -1 \\ 1 & 1 & 1 & \dots & 1 & 1 \end{pmatrix} \xrightarrow{L_n \leftarrow L_n - \frac{1}{1} L_1} \begin{pmatrix} 1 & -1 & \dots & \dots & \dots & -1 \\ 0 & 1 & \ddots & & & \vdots \\ \vdots & \ddots & 1 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & 1 & -1 \\ 0 & 2 & 2 & \dots & 2 & 2 \end{pmatrix} \xrightarrow{L_n \leftarrow L_n - \frac{2}{1} L_2} \begin{pmatrix} 1 & -1 & \dots & \dots & \dots & -1 \\ 0 & 1 & \ddots & & & \vdots \\ \vdots & \ddots & 1 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & 1 & -1 \\ 0 & 0 & 4 & \dots & 4 & 4 \end{pmatrix}$$

$$[\dots] \xrightarrow{L_n \leftarrow L_n - \frac{2^{n-2}}{1} L_{n-1}} \begin{pmatrix} 1 & -1 & \dots & \dots & \dots & -1 \\ 0 & 1 & \ddots & & & \vdots \\ \vdots & \ddots & 1 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & 1 & -1 \\ 0 & 0 & 0 & \dots & 0 & 2^{n-1} \end{pmatrix}.$$

On obtient les matrices

$$\mathbb{L} = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & 0 \\ 0 & 1 & \ddots & & & \vdots \\ 0 & 0 & 1 & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 1 & 0 \\ 1 & 2 & 4 & \dots & 2^{n-2} & 1 \end{pmatrix} \quad \text{et} \quad \mathbb{U} = \begin{pmatrix} 1 & -1 & \dots & \dots & \dots & -1 \\ 0 & 1 & \ddots & & & \vdots \\ 0 & 0 & 1 & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & -1 & \vdots \\ 0 & 0 & \dots & 0 & 1 & -1 \\ 0 & 0 & 0 & \dots & 0 & 2^{n-1} \end{pmatrix}.$$

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,...,n-1, j=1,...,n,
- $u_{nj} = 0$ si $j < n$,
- $u_{nn} = 2^{n-1}$.

Exercice 2.6

Considérons une matrice $\mathbb{A} \in \mathbb{R}^{n\times n}$ (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 $\mathbb{LU}$ de $\mathbb{A}$.

Correction
Factorisation $\mathbb{LU}$ de la matrice $\mathbb{A}$:

$$\begin{pmatrix} 1 & 0 & \dots & \dots & 0 & 1 \\ -1 & 1 & \ddots & & \vdots & \vdots \\ \vdots & \ddots & 1 & \ddots & \vdots & \vdots \\ \vdots & & \ddots & \ddots & 0 & \vdots \\ \vdots & & & \ddots & 1 & 1 \\ -1 & \dots & \dots & \dots & -1 & 1 \end{pmatrix} \xrightarrow[L_n \leftarrow L_n + L_1]{L_2 \leftarrow L_2 + L_1} \begin{pmatrix} 1 & 0 & \dots & \dots & 0 & 1 \\ 0 & 1 & \ddots & & \vdots & 2 \\ \vdots & -1 & 1 & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 & \vdots \\ \vdots & \vdots & & \ddots & 1 & 2 \\ 0 & -1 & \dots & \dots & -1 & 2 \end{pmatrix} \xrightarrow[L_n \leftarrow L_n + L_2]{L_3 \leftarrow L_3 + L_2} \begin{pmatrix} 1 & 0 & \dots & \dots & 0 & 1 \\ 0 & 1 & \ddots & & \vdots & 2 \\ \vdots & 0 & 1 & \ddots & \vdots & 4 \\ \vdots & \vdots & -1 & \ddots & 0 & \vdots \\ \vdots & \vdots & \vdots & \ddots & 1 & 4 \\ 0 & 0 & -1 & \dots & -1 & 4 \end{pmatrix}$$

$$[\dots] \xrightarrow{L_n \leftarrow L_n + L_{n-1}} \begin{pmatrix} 1 & 0 & \dots & \dots & 0 & 2^0 \\ 0 & 1 & \ddots & & \vdots & 2^1 \\ \vdots & \ddots & 1 & \ddots & \vdots & 2^2 \\ \vdots & & \ddots & \ddots & 0 & \vdots \\ \vdots & & & \ddots & 1 & 2^{n-2} \\ 0 & \dots & \dots & \dots & 0 & 2^{n-1} \end{pmatrix}$$

On obtient les matrices

$$\mathbb{L} = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & 0 \\ -1 & 1 & \ddots & & & \vdots \\ \vdots & \ddots & 1 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & 1 & 0 \\ -1 & \dots & \dots & \dots & -1 & 1 \end{pmatrix} \quad \text{et} \quad \mathbb{U} = \begin{pmatrix} 1 & 0 & \dots & \dots & 0 & 2^0 \\ 0 & 1 & \ddots & & \vdots & 2^1 \\ \vdots & \ddots & 1 & \ddots & \vdots & 2^2 \\ \vdots & & \ddots & \ddots & 0 & \vdots \\ \vdots & & & \ddots & 1 & 2^{n-2} \\ 0 & \dots & \dots & \dots & 0 & 2^{n-1} \end{pmatrix}.$$

i.e.

- $\star$ $\ell_{ii} = 1$ pour $i = 1, \dots, n,$,
- $\star$ $\ell_{ij} = -1$ si $i > j$
- $\star$ $\ell_{ij} = 0$ sinon ;
- $\star$ $u_{ii} = 1$ pour $i = 1, \dots, n-1$,
- $\star$ $u_{in} = 2^{i-1}$ pour $i = 1, \dots, n$,
- $\star$ $u_{ij} = 0$ sinon.

Exercice 2.7
Calculer, lorsqu'il est possible, la factorisation $\mathbb{LU}$ des matrices suivantes :

$$\mathbb{A} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 7 & 8 & 9 \end{pmatrix}, \qquad \mathbb{B} = \begin{pmatrix} 1 & 2 & 3 \\ 7 & 8 & 9 \\ 2 & 4 & 5 \end{pmatrix}.$$

Comment peut-on modifier l'algorithme de factorisation pour pouvoir toujours aboutir à une factorisation $\mathbb{LU}$ lorsque la matrice est inversible ?

Correction
Pour une matrice quelconque $\mathbb{A} \in \mathscr{M}_{n,n}(\mathbb{R})$, la factorisation $\mathbb{LU}$ (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 $\det(\mathbb{A}_2) = a_{11}a_{22} - a12a_{21} = 0$, 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 $\mathbb{LU}$ 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 $\mathbb{LU}$ de la matrice $\mathbb{B}$ est donc

$$\mathbb{L} = \begin{pmatrix} 1 & 0 & 0 \\ 7 & 1 & 0 \\ 2 & 0 & 1 \end{pmatrix}, \qquad \mathbb{U} = \begin{pmatrix} 1 & 2 & 3 \\ 0 & -6 & -12 \\ 0 & 0 & -1 \end{pmatrix}.$$

Lorsqu'un pivot est nul, la méthode de GAUSS pour calculer la factorisation $\mathbb{LU}$ 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 $\mathbb{PA} = \mathbb{LU}$: 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 2.8

Soit α un paramètre réel et soient les matrices $\mathbb{A}_\alpha$, $\mathbb{P}$ et le vecteur $\mathbf{b}$ définis par

$$\mathbb{A}_\alpha = \begin{pmatrix} 2 & 4 & 1 \\ \alpha & -2 & -1 \\ 2 & 3 & 2 \end{pmatrix}, \qquad \mathbb{P} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 0 \\ -3/2 \\ -1 \end{pmatrix}.$$

1. À quelle condition sur α, la matrice $\mathbb{A}_\alpha$ est inversible ?
2. À quelle condition sur α, la matrice $\mathbb{A}_\alpha$ admet-elle une décomposition $\mathbb{LU}$ (sans pivot) ?
3. Soit $\alpha = -1$. Calculer, si elle existe, la décomposition $\mathbb{LU}$ de la matrice $\mathbb{M} = \mathbb{PA}_\alpha$.
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 $i = 1, \ldots, n-1$ (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 :

$$\mathbb{A}_1 = \begin{pmatrix} 2 \end{pmatrix}, \qquad \det(\mathbb{A}_1) = 2;$$
$$\mathbb{A}_2 = \begin{pmatrix} 2 & 4 \\ \alpha & -2 \end{pmatrix}, \qquad \det(\mathbb{A}_2) = -4(1+\alpha).$$

Par conséquent, la matrice $\mathbb{A}_\alpha$ admet une décomposition $\mathbb{LU}$ (sans pivot) si et seulement si $\alpha \neq -1$.

3. Si $\alpha = -1$ la matrice $\mathbb{A}_\alpha$ n'admet pas de décomposition $\mathbb{LU}$ sans pivot. La matrice $\mathbb{P}$ échange les lignes 2 et 3 de la matrice $\mathbb{A}$ et on obtient la matrice

$$\mathbb{PA}_{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 2 & 4 & 1 \\ -1 & -2 & -1 \\ 2 & 3 & 2 \end{pmatrix} = \begin{pmatrix} 2 & 4 & 1 \\ 2 & 3 & 2 \\ -1 & -2 & -1 \end{pmatrix}.$$

La matrice $\mathbb{M}$ admet une décomposition $\mathbb{LU}$ (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 $\mathbb{LU}$ suivante de la matrice $\mathbb{M}$:

$$\mathbb{L} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ -\frac{1}{2} & 0 & 1 \end{pmatrix}, \qquad \mathbb{U} = \begin{pmatrix} 2 & 4 & 1 \\ 0 & -1 & 1 \\ 0 & 0 & -\frac{1}{2} \end{pmatrix}.$$

4. Pour résoudre le système linéaire $\mathbb{M}\mathbf{x} = \mathbb{P}\mathbf{b}$ il suffit de résoudre les deux systèmes triangulaires suivantes :

   ⋆ $\mathbb{L}\mathbf{y} = \mathbb{P}\mathbf{b}$:

$$y_1 = 0, \qquad y_2 = -1 - y_1 = -1, \qquad y_3 = -\frac{3}{2} + \frac{1}{2} y_1 = -\frac{3}{2};$$

   ⋆ $\mathbb{U}\mathbf{x} = \mathbf{y}$:

$$x_3 = \frac{-3}{2}(-2) = 3, \qquad x_2 = (-1 - x_3)/(-1) = 4, \qquad x_1 = (0 - 4x_2 - x_3)/2 = -\frac{19}{2}.$$

Exercice 2.9

Considérons les deux matrices carrées d'ordre $n > 3$:

$$\mathbb{A} = \begin{pmatrix} \alpha & 0 & 0 & 0 & & \dots & \beta \\ 0 & \alpha & 0 & 0 & 0 & \dots & \beta \\ 0 & 0 & \alpha & 0 & \ddots & & \vdots \\ & 0 & \ddots & \ddots & & \dots & \beta \\ \vdots & \vdots & & \ddots & & 0 & \beta \\ 0 & 0 & & & 0 & \alpha & \beta \\ \beta & \beta & \dots & & \beta & \beta & \alpha \end{pmatrix} \qquad \mathbb{B} = \begin{pmatrix} \beta & 0 & \dots & \dots & 0 & 0 & \alpha \\ \beta & & 0 & 0 & 0 & \alpha & 0 \\ \vdots & & & 0 & \ddots & & 0 \\ & & & \ddots & & \dots & \vdots \\ \vdots & 0 & \alpha & 0 & & 0 & 0 \\ \beta & \alpha & 0 & & 0 & \alpha & 0 \\ \alpha & \beta & \beta & \dots & & \beta & \beta \end{pmatrix}$$

avec α et β réels non nuls.

1. Vérifier que la factorisation $\mathbb{LU}$ 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 $\mathbb{LU}$ 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 α et β 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 $\mathbb{LU}$ 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 :

$$\mathbb{B}^{(1)} = \begin{pmatrix} \beta & 0 & 0 & \alpha \\ \beta & 0 & \alpha & 0 \\ \beta & \alpha & 0 & 0 \\ \alpha & \beta & \beta & \beta \end{pmatrix} \xrightarrow[L_4 \leftarrow L_4 - \frac{\alpha}{\beta} L_1]{\substack{L_2 \leftarrow L_2 - L_1 \\ L_3 \leftarrow L_3 - L_1}} \mathbb{B}^{(2)} = \begin{pmatrix} \beta & 0 & 0 & \alpha \\ 0 & \boxed{0} & \alpha & -\alpha \\ 0 & \alpha & 0 & -\alpha \\ 0 & \beta & \beta & \beta - \frac{\alpha^2}{\beta} \end{pmatrix}.$$

2. La matrice $\mathbb{A}$ est une matrice «en flèche» : pour en calculer la factorisation $\mathbb{LU}$ 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 globale est donc de l'ordre de n.
3. Le déterminant δ_n de la matrice $\mathbb{A}$ de dimension n coïncide avec le déterminant de la matrice $\mathbb{U}$. 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 précédant, ce déterminant s'écrit

$$(\alpha - \lambda)^n - (n-1)(\alpha - \lambda)^{n-2}\beta^2$$

dont les racines sont

$$\lambda_{1,2} = \alpha \pm \sqrt{(n-1)\beta}, \qquad \lambda_3 = \cdots = \lambda_n = \alpha.$$

Par conséquent, pour que la matrice $\mathbb{A}$ soit définie positive il faut que les valeurs propres soient tous positifs, ce qui impose

$$\alpha > 0, \qquad |\beta| < \frac{\alpha}{\sqrt{n-1}}.$$

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 2.10

Écrire les formules de la méthode d'élimination de GAUSS pour une matrice de la forme

$$\mathbb{A} = \begin{pmatrix} a_{1,1} & a_{1,2} & 0 & \dots & & 0 \\ a_{2,1} & a_{2,2} & a_{2,3} & 0 & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & 0 \\ \vdots & & & & a_{n-1,n-1} & a_{n-1,n} \\ a_{n,1} & a_{n,2} & \dots & & a_{n,n-1} & a_{n,n} \end{pmatrix}.$$

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

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} + \ell_{ik} a_{k,j}^{(k)}, \qquad i, j = k+1,$$
$$\ell_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}, \qquad i = k+1.$$

La coût est donc de l'ordre de n et la matrice $\mathbb{U}$ est bidiagonale supérieure.

Exercice 2.11

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 $1/\varepsilon$ (ε 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 $\mathbb{LU}$ des deux matrices associées à ces systèmes. On fournira également la solution exacte de ces systèmes. Commenter.

Correction

Il s'agit du même système linéaire (on a juste échangé l'ordre des équations) donc la solution exacte est la même. Pour un système 2×2 il n'est même pas nécessaire d'utiliser la méthode de Gauss, on peut directement calculer la solution car

$$\begin{cases} \varepsilon a + b = 1 \\ 2a + b = 0 \end{cases} \iff \begin{cases} \varepsilon a - 2a = 1 \\ b = -2a \end{cases} \iff \begin{cases} (\varepsilon - 2)a = 1 \\ b = -2a \end{cases}$$

Si $\varepsilon = 2$ il n'y a pas de solution ; si $\varepsilon \neq 2$ alors $a = \dfrac{1}{\varepsilon - 2}$ et $b = \dfrac{2}{2-\varepsilon}$ donc, si $\varepsilon \approx 0$, on a $a \simeq -\dfrac{1}{2}$ et $b \simeq 1$.

Premier système :

$$\begin{pmatrix} \varepsilon & 1 \\ 2 & 1 \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Factorisation $\mathbb{LU}$:

$$\begin{pmatrix} \varepsilon & 1 \\ 2 & 1 \end{pmatrix} \xrightarrow{L_2 \leftarrow L_2 - \frac{2}{\varepsilon} L_1} \begin{pmatrix} \varepsilon & 1 \\ 0 & 1 - \frac{2}{\varepsilon} \end{pmatrix} \quad \text{donc} \quad \mathbb{L} = \begin{pmatrix} 1 & 0 \\ \frac{2}{\varepsilon} & 1 \end{pmatrix}, \quad \mathbb{U} = \begin{pmatrix} \varepsilon & 1 \\ 0 & 1 - \frac{2}{\varepsilon} \end{pmatrix}$$

Pour résoudre le système linéaire on résout les systèmes triangulaires $\mathbb{L}\mathbf{y} = \mathbf{b}$ et $\mathbb{U}\mathbf{x} = \mathbf{y}$:

$$\begin{pmatrix} 1 & 0 \\ \frac{2}{\varepsilon} & 1 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \implies y_1 = 1, \quad y_2 = -\frac{2}{\varepsilon};$$

$$\begin{pmatrix} \varepsilon & 1 \\ 0 & 1 - \frac{2}{\varepsilon} \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 1 \\ -\frac{2}{\varepsilon} \end{pmatrix} \implies b = -\frac{2}{\varepsilon\left(1 - \frac{2}{\varepsilon}\right)} = \frac{2}{2-\varepsilon}, \quad a = \frac{1-b}{\varepsilon} = \frac{1}{\varepsilon - 2}.$$

On retrouve bien la solution calculée. Cependant avec l'ordinateur, comme $1+(1/\varepsilon) \approx 1/\varepsilon$, on obtient la même matrice $\mathbb{L}$ mais juste une approximation de la matrice $\mathbb{U}$:

$$\widetilde{\mathbb{L}} = \mathbb{L} = \begin{pmatrix} 1 & 0 \\ \frac{2}{\varepsilon} & 1 \end{pmatrix} \qquad \widetilde{\mathbb{U}} = \begin{pmatrix} \varepsilon & 1 \\ 0 & -\frac{2}{\varepsilon} \end{pmatrix}$$

Pour résoudre ce système linéaire approché on résout les systèmes triangulaires $\widetilde{\mathbb{L}}\mathbf{y} = \mathbf{b}$ et $\widetilde{\mathbb{U}}\mathbf{x} = \mathbf{y}$:

$$\begin{pmatrix} 1 & 0 \\ \frac{2}{\varepsilon} & 1 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \implies y_1 = 1, \quad y_2 = -\frac{2}{\varepsilon};$$

$$\begin{pmatrix} \varepsilon & 1 \\ 0 & -\frac{2}{\varepsilon} \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 1 \\ -\frac{2}{\varepsilon} \end{pmatrix} \implies b = 1, \quad a = 0,$$

ce qui est bien différent de la solution exacte $a \simeq -\frac{1}{2}$.

Second système :

$$\begin{pmatrix} 2 & 1 \\ \varepsilon & 1 \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Factorisation $\mathbb{LU}$:

$$\begin{pmatrix} 2 & 1 \\ \varepsilon & 1 \end{pmatrix} \xrightarrow{L_2 \leftarrow L_2 - \frac{\varepsilon}{2} L_1} \begin{pmatrix} 2 & 1 \\ 0 & 1 - \frac{\varepsilon}{2} \end{pmatrix} \quad \text{donc} \quad \mathbb{L} = \begin{pmatrix} 1 & 0 \\ \frac{\varepsilon}{2} & 1 \end{pmatrix}, \quad \mathbb{U} = \begin{pmatrix} 2 & 1 \\ 0 & 1 - \frac{\varepsilon}{2} \end{pmatrix}$$

Pour résoudre le système linéaire on résout les systèmes triangulaires $\mathbb{L}\mathbf{y} = \mathbf{b}$ et $\mathbb{U}\mathbf{x} = \mathbf{y}$:

$$\begin{pmatrix} 1 & 0 \\ \frac{\varepsilon}{2} & 1 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \implies y_1 = 0, \quad y_2 = 1;$$

$$\begin{pmatrix}2 & 1\\ 0 & 1-\frac{\varepsilon}{2}\end{pmatrix}\begin{pmatrix}a\\ b\end{pmatrix}=\begin{pmatrix}0\\ 1\end{pmatrix} \qquad \Longrightarrow \qquad b=\frac{1}{1-\frac{\varepsilon}{2}}=\frac{2}{2-\varepsilon},\quad a=\frac{0-b}{2}=\frac{1}{\varepsilon-2}.$$

on retrouve bien la solution exacte. Cependant avec l'ordinateur, comme $1+(1/\varepsilon)\approx 1/\varepsilon$, on obtient la même matrice $\mathbb{L}$ mais juste une approximation de la matrice $\mathbb{U}$:

$$\widetilde{\mathbb{L}}=\mathbb{L}=\begin{pmatrix}1 & 0\\ \frac{\varepsilon}{2} & 1\end{pmatrix} \qquad\qquad \widetilde{\mathbb{U}}=\begin{pmatrix}2 & 1\\ 0 & -\frac{2}{\varepsilon}\end{pmatrix}$$

Pour résoudre ce système linéaire approché on résout les systèmes triangulaires $\widetilde{\mathbb{L}}\mathbf{y}=\mathbf{b}$ et $\widetilde{\mathbb{U}}\mathbf{x}=\mathbf{y}$:

$$\begin{pmatrix}1 & 0\\ \frac{\varepsilon}{2} & 1\end{pmatrix}\begin{pmatrix}y_1\\ y_2\end{pmatrix}=\begin{pmatrix}0\\ 1\end{pmatrix} \qquad \Longrightarrow \qquad y_1=0,\quad y_2=1;$$
$$\begin{pmatrix}2 & 1\\ 0 & -\frac{2}{\varepsilon}\end{pmatrix}\begin{pmatrix}a\\ b\end{pmatrix}=\begin{pmatrix}0\\ 1\end{pmatrix} \qquad \Longrightarrow \qquad b=-\frac{\varepsilon}{2},\quad a=\frac{\varepsilon}{4}.$$

Pour $\varepsilon\approx 0$ on obtient $a\simeq 0$ et $b\simeq 0$, ce qui est loin de la solution exacte $a\simeq -\frac{1}{2}$ et $b\simeq 1$.

Exercice 2.12

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 :

$$\left(\begin{array}{ccc|c}1 & 1 & 1 & 1\\ 2 & 1 & 3 & 1\\ -3 & 2 & 4 & 1\end{array}\right)\xrightarrow[L_3\leftarrow L_3-\frac{-3}{1}L_1]{L_2\leftarrow L_2-\frac{2}{1}L_1}\left(\begin{array}{ccc|c}1 & 1 & 1 & 1\\ 0 & -1 & 1 & -1\\ 0 & 5 & 7 & 4\end{array}\right)\xrightarrow{L_3\leftarrow L_3-\frac{5}{-1}L_2}\left(\begin{array}{ccc|c}1 & 1 & 1 & 1\\ 0 & -1 & 1 & -1\\ 0 & 0 & 12 & -1\end{array}\right)$$

donc

$$\mathbb{L}=\begin{pmatrix}1 & 0 & 0\\ 2 & 1 & 0\\ -3 & -5 & 1\end{pmatrix} \qquad \mathbb{U}=\begin{pmatrix}1 & 1 & 1\\ 0 & -1 & 1\\ 0 & 0 & 12\end{pmatrix}$$

Il ne reste à résoudre que le système triangulaire

$$\begin{cases}x_1+x_2+x_3=1\\ -x_2+x_3=-1\\ 12x_3=-1\end{cases} \Longrightarrow x_3=-\frac{1}{12},\quad x_2=\frac{11}{12},\quad x_1=\frac{1}{6}.$$

Deuxième système :

$$\left(\begin{array}{cccc|c}1 & 2 & 3 & 4 & 1\\ 2 & -5 & 7 & 1 & 1\\ 3 & 1 & 1 & 5 & 1\\ 2 & 2 & 0 & 3 & 1\end{array}\right)\xrightarrow[L_4\leftarrow L_4-\frac{2}{1}L_1]{\substack{L_2\leftarrow L_2-\frac{2}{1}L_1\\ L_3\leftarrow L_3-\frac{3}{1}L_1}}\left(\begin{array}{cccc|c}1 & 2 & 3 & 4 & 1\\ 0 & -9 & 1 & -7 & -1\\ 0 & -5 & -8 & -7 & -2\\ 0 & -2 & -6 & -5 & -1\end{array}\right)$$

$$\xrightarrow[L_4\leftarrow L_4-\frac{-2}{-9}L_2]{L_3\leftarrow L_3-\frac{-5}{-9}L_2}\left(\begin{array}{cccc|c}1 & 2 & 3 & 4 & 1\\ 0 & -9 & 1 & -7 & -1\\ 0 & 0 & -\frac{77}{9} & -\frac{28}{9} & -\frac{13}{9}\\ 0 & 0 & -\frac{56}{9} & -\frac{31}{9} & -\frac{7}{9}\end{array}\right)\xrightarrow{L_4\leftarrow L_4-\frac{56/9}{77/9}L_2}\left(\begin{array}{cccc|c}1 & 2 & 3 & 4 & 1\\ 0 & -9 & 1 & -7 & -1\\ 0 & 0 & -\frac{77}{9} & -\frac{28}{9} & -\frac{13}{9}\\ 0 & 0 & 0 & -\frac{13}{11} & \frac{3}{11}\end{array}\right)$$

donc

$$\mathbb{L} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & \frac{5}{9} & 1 & 0 \\ 2 & \frac{2}{9} & \frac{56}{77} & 1 \end{pmatrix} \qquad \mathbb{U} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & -9 & 1 & -7 \\ 0 & 0 & -\frac{77}{9} & -\frac{28}{9} \\ 0 & 0 & 0 & -\frac{13}{11} \end{pmatrix}$$

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 :

$$\left(\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 1 \\ 1 & -2 & 3 & 4 & 1 \\ 1 & 4 & 6 & 8 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{array}\right) \xrightarrow[L_4 \leftarrow L_4 - L_1]{L_2 \leftarrow L_2 - L_1 \\ L_3 \leftarrow L_3 - L_1} \left(\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 1 \\ 0 & -3 & 2 & 3 & 0 \\ 0 & 3 & 5 & 7 & 0 \\ 0 & -1 & -1 & -1 & 0 \end{array}\right)$$

$$\xrightarrow[L_4 \leftarrow L_4 - \frac{-1}{-3} L_2]{L_3 \leftarrow L_3 - (-1) L_2} \left(\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 1 \\ 0 & -3 & 2 & 3 & -0 \\ 0 & 0 & 7 & 10 & 0 \\ 0 & 0 & -\frac{5}{3} & -2 & 0 \end{array}\right) \xrightarrow{L_4 \leftarrow L_4 - \frac{-5/3}{7} L_2} \left(\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 1 \\ 0 & -3 & 2 & 3 & -0 \\ 0 & 0 & 7 & 10 & 0 \\ 0 & 0 & 0 & \frac{8}{21} & 0 \end{array}\right)$$

donc

$$\mathbb{L} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & -1 & 1 & 0 \\ 1 & \frac{1}{3} & \frac{-5}{21} & 1 \end{pmatrix} \qquad \mathbb{U} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & -3 & 2 & 3 \\ 0 & 0 & 7 & 10 \\ 0 & 0 & 0 & \frac{8}{21} \end{pmatrix}$$

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 $\mathbb{QR}$ et systèmes linéaires sur déterminés

☆ Exercice 2.13 (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 :

$$\mathbb{A} = \begin{pmatrix} 0 & 1 \\ 0.06 & 1 \\ 0.14 & 1 \\ 0.25 & 1 \\ 0.31 & 1 \\ 0.47 & 1 \\ 0.6 & 1 \\ 0.7 & 1 \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} 0 \\ 0.08 \\ 0.14 \\ 0.2 \\ 0.23 \\ 0.25 \\ 0.28 \\ 0.29 \end{pmatrix}$$

Calculer la solution $\mathbf{x} \in \mathbb{R}^2$ au sens des moindres carrés en utilisant la factorisation $\mathbb{QR}$. 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 `A\b`.

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 2.14 (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 $\mathbf{x}^{(k)} = (x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)})^T \in \mathbb{R}^n$ convergent vers le vecteur solution exacte $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ pour tout vecteur initiale $\mathbf{x}^{(0)} = (x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)})^T \in \mathbb{R}^n$ lorsque k tend vers $+\infty$.

Méthode de Jacobi Soit $\mathbf{x}^0 = (x_1^0, x_2^0, \ldots, x_n^0)$ 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 :

$$x_i^{k+1} = \frac{b_i - \sum\limits_{\substack{j=1 \\ j\neq i}}^{n} a_{ij} x_j^k}{a_{ii}}, \qquad i = 1, \ldots, n$$

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 $\mathbf{x}^0 = (x_1^0, x_2^0, \ldots, x_n^0)$ 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 :

$$x_i^{k+1} = \frac{b_i - \sum\limits_{j=1}^{i-1} a_{ij} x_j^{k+1} - \sum\limits_{j=i+1}^{n} a_{ij} x_j^k}{a_{ii}}, \qquad i = 1, \ldots, n$$

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 `function [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 `x` un vecteur colonne de $\mathbb{R}^n$ solution du système linaire $\mathbb{A}\mathbf{x} = \mathbf{b}$, `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`$= 50$,

$$\mathbb{A} = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad \mathbf{x}^{(0)} = \begin{pmatrix} 1 \\ 1/2 \end{pmatrix}.$$

   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
         j=[1:i-1,i+1:n];
         x(i)=(b(i)-dot(A(i,j),xold(j)))/A(i,i);
      end
      k+=1;
      xold=x';
      r=norm(A*xold-b);
      disp(k)
      disp(x)
   end
end

% Remarque: sum(A(i,j).*xold(j))=dot(A(i,j),xold(j))
```

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)
%A\b % verification avec la fonction predefinie dans Octave
%
%E=-tril(A,-1);
%F=-triu(A,1);
%P=A+E+F;
%B=inv(P)*(P-A);
%RayonSpectraleB_Jacobi=max(abs(eig(B)))

clear all;
clc;
A=[1 2 4; 2 1 6; 4 6 1]
b=[8;3;7]
xinit=[0;0;0];
disp("Jacobi")
[x,r,k]=myJacobi(A,b,xinit,1.e-9,2)
A\b % verification avec la fonction predefinie dans Octave

E=-tril(A,-1);
F=-triu(A,1);
P=A+E+F;
B=inv(P)*(P-A);
RayonSpectraleB_Jacobi=max(abs(eig(B)))


disp("Gauss Seidel")
[x,r,k]=myGS(A,b,xinit,1.e-9,2)
A\b % verification avec la fonction predefinie dans Octave

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)))
```

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
         j=[1:i-1,i+1:n];
```

```
        x(i)=(b(i)-dot(A(i,j),x(j)))/A(i,i);
      end
      k+=1;
      r=norm(A*x-b);
      disp(k)
      disp(x)
    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)
A\b % verification avec la fonction predefinie dans Octave

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 2.15

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 $\mathbb{A}$ (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 :

$$\mathbf{x}^{(0)} = \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix}, \quad \mathbf{x}^{(1)} = \begin{pmatrix} \frac{12-(1\times 2+1\times 2)}{6} \\ \frac{0-(2\times \frac{4}{3}+0\times 2)}{4} \\ \frac{6-(1\times \frac{4}{3}+2\times \frac{-2}{3})}{6} \end{pmatrix} = \begin{pmatrix} {}^{4}/_{3} \\ -\frac{2}{3} \\ 1 \end{pmatrix}, \quad \mathbf{x}^{(2)} = \begin{pmatrix} \frac{12-(1\times \frac{-2}{3}+1\times 1)}{6} \\ \frac{0-(2\times \frac{35}{18}+0\times 1)}{4} \\ \frac{6-(1\times \frac{35}{18}+2\times \frac{-35}{36})}{6} \end{pmatrix} = \begin{pmatrix} {}^{35}/_{18} \\ -{}^{35}/_{36} \\ 1 \end{pmatrix}, \quad \mathbf{x}^{(3)} = \begin{pmatrix} \frac{12-(1\times \frac{35}{18}+1\times \frac{-35}{36})}{6} \\ \frac{0-(2\times \frac{431}{216}+0\times 1)}{4} \\ \frac{6-(1\times \frac{431}{216}+2\times \frac{-431}{432})}{6} \end{pmatrix} = \begin{pmatrix} {}^{431}/_{216} \\ -{}^{431}/_{432} \\ 1 \end{pmatrix}$$

ainsi

$$\mathbf{x} \approx \begin{pmatrix} 1.995 \\ -0.995 \\ 1 \end{pmatrix}.$$

3. Méthode d'élimination de GAUSS :

$$(\mathbb{A}|\mathbf{b}) = \left(\begin{array}{ccc|c} 6 & 1 & 1 & 12 \\ 2 & 4 & 0 & 0 \\ 1 & 2 & 6 & 6 \end{array}\right) \xrightarrow[L_3 \leftarrow L_3 - \frac{1}{6}L_1]{L_2 \leftarrow L_2 - \frac{2}{6}L_1} \left(\begin{array}{ccc|c} 6 & 1 & 1 & 12 \\ 0 & \frac{11}{3} & -\frac{1}{3} & -4 \\ 0 & \frac{11}{6} & \frac{35}{6} & 4 \end{array}\right) \xrightarrow{L_3 \leftarrow L_3 - \frac{\frac{11}{6}}{\frac{11}{3}}L_2} \left(\begin{array}{ccc|c} 6 & 1 & 1 & 12 \\ 0 & \frac{11}{3} & -\frac{1}{3} & -4 \\ 0 & 0 & 6 & 6 \end{array}\right)$$

donc

$$\begin{cases} 6x_1 + x_2 + x_3 = 12, \\ \frac{11}{3}x_2 - \frac{1}{3}x_3 = -4 \\ 6x_3 = 6 \end{cases} \quad \Longrightarrow \quad x_3 = 1, \quad x_2 = -1, \quad x_1 = 2.$$

4. Factorisation de la matrice $\mathbb{A}$:

$$\begin{pmatrix} 6 & 1 & 1 \\ 2 & 4 & 0 \\ 1 & 2 & 6 \end{pmatrix} \xrightarrow[L_3 \leftarrow L_3 - \frac{1}{6}L_1]{L_2 \leftarrow L_2 - \frac{2}{6}L_1} \begin{pmatrix} 6 & 1 & 1 \\ \frac{2}{6} & \frac{11}{3} & -\frac{1}{3} \\ \frac{1}{6} & \frac{11}{6} & \frac{35}{6} \end{pmatrix} \xrightarrow{L_3 \leftarrow L_3 - \frac{\frac{11}{6}}{\frac{11}{3}} L_2} \begin{pmatrix} 6 & 1 & 1 \\ \frac{2}{6} & \frac{11}{3} & -\frac{1}{3} \\ \frac{1}{6} & \frac{11}{6} & 6 \end{pmatrix}$$

donc

$$\mathbb{L} = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{3} & 1 & 0 \\ \frac{1}{6} & \frac{1}{2} & 1 \end{pmatrix} \qquad \mathbb{U} = \begin{pmatrix} 6 & 1 & 1 \\ 0 & \frac{11}{3} & -\frac{1}{3} \\ 0 & 0 & 6 \end{pmatrix}$$

Pour résoudre le système linéaire on résout les systèmes triangulaires $\mathbb{L}\mathbf{y} = \mathbf{b}$

$$\begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{3} & 1 & 0 \\ \frac{1}{6} & \frac{1}{2} & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 12 \\ 0 \\ 6 \end{pmatrix} \quad \Longrightarrow \quad y_1 = 12, \quad y_2 = -4, \quad y_3 = 6$$

et $\mathbb{U}\mathbf{x} = \mathbf{y}$

$$\begin{pmatrix} 6 & 1 & 1 \\ 0 & \frac{11}{3} & -\frac{1}{3} \\ 0 & 0 & 6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ -4 \\ 6 \end{pmatrix} \quad \Longrightarrow \quad x_3 = 1, \quad x_2 = -1, \quad x_1 = 2.$$

Exercice 2.16

Donner une condition suffisante sur le coefficient α 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 2.17

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 α, β, γ et δ 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
$$\lambda_1 = \alpha, \qquad \lambda_2 = \alpha - \beta, \qquad \lambda_3 = \alpha + \beta,$$
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 2.18

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 :
$$\left(\begin{array}{cc|c} 10 & 1 & 11 \\ 2 & 10 & 12 \end{array}\right) \xrightarrow{L_2 \leftarrow L_2 - \frac{2}{10} L_1} \left(\begin{array}{cc|c} 10 & 1 & 11 \\ 0 & \frac{49}{5} & \frac{49}{5} \end{array}\right) \Longrightarrow \begin{cases} 10a + b = 11 \\ \frac{49}{5} b = \frac{49}{5} \end{cases} \Longrightarrow \begin{cases} a = 1 \\ b = 1. \end{cases}$$

⋆ Second système :
$$\left(\begin{array}{cc|c} 2 & 10 & 12 \\ 10 & 1 & 11 \end{array}\right) \xrightarrow{L_2 \leftarrow L_2 - \frac{10}{2} L_1} \left(\begin{array}{cc|c} 2 & 10 & 12 \\ 0 & -49 & -49 \end{array}\right) \Longrightarrow \begin{cases} 2a + 10b = 12 \\ -49b = -49 \end{cases} \Longrightarrow \begin{cases} a = 1 \\ b = 1. \end{cases}$$

Jacobi ⋆ Premier système :
$$\begin{cases} 10a + b = 11 \\ 2a + 10b = 12 \end{cases} \Longleftrightarrow \begin{cases} a = \frac{11 - b}{10} \\ b = \frac{12 - 2a}{10} \end{cases}$$
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 :
$$\begin{cases} 2a + 10b = 12 \\ 10a + b = 11 \end{cases} \Longleftrightarrow \begin{cases} a = \frac{12 - 10b}{2} \\ b = 11 - 10a \end{cases}$$
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 :
$$\begin{cases} 10a + b = 11 \\ 2a + 10b = 12 \end{cases} \Longleftrightarrow \begin{cases} a = \frac{11 - b}{10} \\ b = \frac{12 - 2a}{10} \end{cases}$$

La matrice étant à diagonale dominante stricte, la méthode converge et on a

$$\mathbf{x}^{(0)} = \begin{pmatrix}0\\0\end{pmatrix}, \quad \mathbf{x}^{(1)} = \begin{pmatrix}\frac{11-0}{10}\\ \frac{12-2\frac{11}{10}}{10}\end{pmatrix} = \begin{pmatrix}{}^{11}/_{10}\\ {}^{49}/_{50}\end{pmatrix}, \quad \mathbf{x}^{(2)} = \begin{pmatrix}\frac{11-\frac{49}{50}}{10}\\ \frac{12-2\frac{501}{500}}{10}\end{pmatrix} = \begin{pmatrix}{}^{501}/_{500}\\ {}^{2499}/_{2500}\end{pmatrix}, \quad \mathbf{x}^{(3)} = \begin{pmatrix}\frac{11-\frac{2499}{2500}}{10}\\ \frac{12-2\frac{25001}{25000}}{10}\end{pmatrix} = \begin{pmatrix}{}^{25001}/_{25000}\\ {}^{12499}/_{125000}\end{pmatrix}.$$

⋆ 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 2.19

Soit $\mathbb{A}$ une matrice, $\mathbb{A}\in\mathscr{M}_{n,n}(\mathbb{R})$.

1. Rappeler la méthode de JACOBI pour la résolution du système $\mathbb{A}\mathbf{x}=\mathbf{b}$, avec $\mathbf{b}\in\mathscr{M}_{n,1}(\mathbb{R})$ 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 $\mathbb{LU}$ 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 $\mathbf{x}^{(0)}\in\mathbb{R}^n$:
$$x_i^{k+1} = \frac{b_i - \sum\limits_{\substack{j=1\\ j\neq i}}^{n} a_{ij}x_j^k}{a_{ii}}, \qquad i=1,\dots,n.$$
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 :
$$\begin{pmatrix}4&-1&-1\\-1&3&-1\\-1&-1&4\end{pmatrix} \xrightarrow[L_3\leftarrow L_3-\frac{-1}{4}L_1]{L_2\leftarrow L_2-\frac{-1}{4}L_1} \begin{pmatrix}4&-1&-1\\0&{}^{11}/_4&-{}^5/_4\\0&-{}^5/_4&{}^{15}/_4\end{pmatrix} \xrightarrow{L_3\leftarrow L_3-\frac{-5/4}{11/4}L_2} \begin{pmatrix}4&-1&-1\\0&{}^{11}/_4&-{}^5/_4\\0&0&{}^{35}/_{11}\end{pmatrix}.$$
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 2.20

Soit les systèmes linéaires

$$\begin{cases}4x_1+3x_2+3x_3=10\\ 3x_1+4x_2+3x_3=10\\ 3x_1+3x_2+4x_3=10\end{cases} \tag{2.1}$$

$$\begin{cases}4x_1+x_2+x_3=6\\ x_1+4x_2+x_3=6\\ x_1+x_2+4x_3=6\end{cases} \tag{2.2}$$

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 (2.1) et (2.2) 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 (2.1),
   3.2. avec la méthode de GAUSS-SEIDEL pour le système (2.1),
   3.3. avec la méthode de JACOBI pour le système (2.2),
   3.4. avec la méthode de GAUSS-SEIDEL pour le système (2.2).
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[4] 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

	$\mathbb{A}_1\mathbf{x} = \mathbf{b}$	$\mathbb{A}_2\mathbf{x} = \mathbf{b}$
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 (2.1) :

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(0)} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \Longrightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(1)} = \frac{1}{4} \begin{pmatrix} 10 - 3\times 0 - 3\times 0 \\ 10 - 3\times 0 - 3\times 0 \\ 10 - 3\times 0 - 3\times 0 \end{pmatrix} = \begin{pmatrix} \frac{5}{2} \\ \frac{5}{2} \\ \frac{5}{2} \end{pmatrix}$$

$$\Longrightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(2)} = \frac{1}{4} \begin{pmatrix} 10 - 3\times \frac{5}{2} - 3\times \frac{5}{2} \\ 10 - 3\times \frac{5}{2} - 3\times \frac{5}{2} \\ 10 - 3\times \frac{5}{2} - 3\times \frac{5}{2} \end{pmatrix} = \begin{pmatrix} -\frac{5}{4} \\ -\frac{5}{4} \\ -\frac{5}{4} \end{pmatrix} \Longrightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(3)} = \frac{1}{4} \begin{pmatrix} 10 - 3\times \frac{-5}{4} - 3\times \frac{-5}{4} \\ 10 - 3\times \frac{-5}{4} - 3\times \frac{-5}{4} \\ 10 - 3\times \frac{-5}{4} - 3\times \frac{-5}{4} \end{pmatrix} = \begin{pmatrix} \frac{35}{8} \\ \frac{35}{8} \\ \frac{35}{8} \end{pmatrix}$$

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 $\det \mathbb{A}_1(\lambda) = (\lambda - 1)(-\lambda^2 + 11\lambda - 10) = (\lambda - 1)^2(\lambda - 10)$.

3.2. GAUSS-SEIDEL pour le système (2.1) :

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(0)} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \Longrightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(1)} = \frac{1}{4}\begin{pmatrix} 10-3\times 0-3\times 0 \\ 10-3\times \frac{5}{2}-3\times 0 \\ 10-3\times \frac{5}{2}-3\times \frac{5}{8} \end{pmatrix} = \begin{pmatrix} \frac{5}{2} \\ \frac{5}{8} \\ \frac{5}{32} \end{pmatrix}$$

$$\Longrightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(2)} = \frac{1}{4}\begin{pmatrix} 10-3\times \frac{5}{8}-3\times \frac{5}{32} \\ 10-3\times \frac{245}{128}-3\times \frac{5}{32} \\ 10-3\times \frac{245}{128}-3\times \frac{485}{512} \end{pmatrix} = \begin{pmatrix} \frac{245}{128} \\ \frac{485}{512} \\ \frac{725}{2048} \end{pmatrix} \Longrightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(3)} = \begin{pmatrix} \frac{12485}{8192} \\ \frac{35765}{32768} \\ \frac{70565}{131072} \end{pmatrix}$$

3.3. JACOBI pour le système (2.2) :

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(0)} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \Longrightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(1)} = \frac{1}{4}\begin{pmatrix} 6-1\times 0-1\times 0 \\ 6-1\times 0-1\times 0 \\ 6-1\times 0-1\times 0 \end{pmatrix} = \begin{pmatrix} \frac{3}{2} \\ \frac{3}{2} \\ \frac{3}{2} \end{pmatrix}$$

$$\Longrightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(2)} = \frac{1}{4}\begin{pmatrix} 6-1\times \frac{3}{2}-1\times \frac{3}{2} \\ 6-1\times \frac{3}{2}-1\times \frac{3}{2} \\ 6-1\times \frac{3}{2}-1\times \frac{3}{2} \end{pmatrix} = \begin{pmatrix} \frac{3}{4} \\ \frac{3}{4} \\ \frac{3}{4} \end{pmatrix} \Longrightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(3)} = \frac{1}{4}\begin{pmatrix} 6-1\times \frac{3}{4}-1\times \frac{3}{4} \\ 6-1\times \frac{3}{4}-1\times \frac{3}{4} \\ 6-1\times \frac{3}{4}-1\times \frac{3}{4} \end{pmatrix} = \begin{pmatrix} \frac{9}{8} \\ \frac{9}{8} \\ \frac{9}{8} \end{pmatrix}$$

3.4. GAUSS-SEIDEL pour le système (2.2) :

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(0)} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \Longrightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(1)} = \frac{1}{4}\begin{pmatrix} 6-1\times 0-1\times 0 \\ 6-1\times \frac{3}{2}-1\times 0 \\ 6-1\times \frac{3}{2}-1\times \frac{9}{8} \end{pmatrix} \begin{pmatrix} \frac{3}{2} \\ \frac{9}{8} \\ \frac{27}{32} \end{pmatrix}$$

$$\Longrightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(2)} = \frac{1}{4}\begin{pmatrix} 6-1\times \frac{9}{8}-1\times \frac{27}{32} \\ 6-1\times \frac{129}{128}-1\times \frac{27}{32} \\ 6-1\times \frac{129}{128}-1\times \frac{531}{512} \end{pmatrix} = \begin{pmatrix} \frac{129}{128} \\ \frac{531}{512} \\ \frac{2025}{2048} \end{pmatrix} \Longrightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(3)} = \frac{1}{4}\begin{pmatrix} 6-1\times \frac{531}{512}-1\times \frac{2025}{2048} \\ 6-1\times \frac{8139}{8192}-1\times \frac{2025}{2048} \\ 6-1\times \frac{8139}{8192}-1\times \frac{32913}{32768} \end{pmatrix} = \begin{pmatrix} \frac{8139}{8192} \\ \frac{32913}{32768} \\ \frac{131139}{131072} \end{pmatrix}$$

4. Calcul de la solution exacte à l'aide de la méthode d'élimination de GAUSS :

⋆ Système (2.1) :

$$\left(\begin{array}{ccc|c} 4 & 3 & 3 & 10 \\ 3 & 4 & 3 & 10 \\ 3 & 3 & 4 & 10 \end{array}\right) \xrightarrow[L_3\leftarrow L_3-\frac{3}{4}L_1]{L_2\leftarrow L_2-\frac{3}{4}L_1} \left(\begin{array}{ccc|c} 4 & 3 & 3 & 10 \\ 0 & 7/4 & 3/4 & 5/2 \\ 0 & 3/4 & 7/4 & 5/2 \end{array}\right) \xrightarrow{L_3\leftarrow L_3-\frac{3/4}{7/4}L_2} \left(\begin{array}{ccc|c} 4 & 3 & 3 & 10 \\ 0 & 7/4 & 3/4 & 5/2 \\ 0 & 0 & 10/7 & 10/7 \end{array}\right) \Longrightarrow \mathbf{x} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

⋆ Système (2.2) :

$$\left(\begin{array}{ccc|c} 4 & 1 & 1 & 6 \\ 1 & 4 & 1 & 6 \\ 1 & 1 & 4 & 6 \end{array}\right) \xrightarrow[L_3\leftarrow L_3-\frac{1}{4}L_1]{L_2\leftarrow L_2-\frac{1}{4}L_1} \left(\begin{array}{ccc|c} 4 & 1 & 1 & 6 \\ 0 & 15/4 & 3/4 & 9/2 \\ 0 & 3/4 & 15/4 & 9/2 \end{array}\right) \xrightarrow{L_3\leftarrow L_3-\frac{3/4}{15/4}L_2} \left(\begin{array}{ccc|c} 4 & 1 & 1 & 6 \\ 0 & 15/4 & 3/4 & 9/2 \\ 0 & 0 & 18/5 & 18/5 \end{array}\right) \Longrightarrow \mathbf{x} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

CHAPITRE 3

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 φ interpole l'ensemble de valeurs $\{y_i\}_{i=0}^{n}$ aux nœuds $\{x_i\}_{i=0}^{n}$. Les quantités y_i représentent les valeurs aux nœuds 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 φ est un polynôme et d'interpolation polynomiale par morceaux (ou d'interpolation par fonctions *splines*) si φ est polynomiale par morceaux.

3.1. 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 $x_0, x_1, x_2, \ldots, x_n$ distinctes données (appelés nœuds d'interpolation), prenne les valeurs $y_0, y_1, y_2, \ldots, y_n$ respectivement, c'est-à-dire

$$p_n(x_i) = y_i \qquad \text{pour } 0 \leq i \leq n. \tag{3.1}$$

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 $\mathbb{R}_n[x]$:

$$p_n(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n,$$

où $a_0, a_1, a_2, \ldots, a_n$ sont des coefficients qui devront être déterminés. Les $(n+1)$ relations (3.1) s'écrivent alors

$$\begin{cases} a_0 + a_1 x_0 + \ldots a_n x_0^n = y_0 \\ a_0 + a_1 x_1 + \ldots a_n x_1^n = y_1 \\ \ldots \\ a_0 + a_1 x_n + \ldots a_n x_n^n = y_n \end{cases}$$

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 $a_0, a_1, a_2, \ldots, a_n$ qu'on peut mettre sous la forme matricielle

$$\begin{pmatrix} 1 & x_0 & \ldots & x_0^n \\ 1 & x_1 & \ldots & x_1^n \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \ldots & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}. \tag{3.2}$$

Ainsi, le problème consistant à chercher le polynôme p_n satisfaisant (3.1) peut se réduire à résoudre le système linéaire (3.2) (cette matrice s'appelle matrice de VANDERMONDE).

Étant donné $n+1$ points distincts $x_0, \ldots, x_n$ et $n+1$ valeurs correspondantes $y_0, \ldots, y_n$, 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

$$p_n(x) = \sum_{i=0}^{n} a_i x^i \qquad \text{avec} \quad \begin{pmatrix} 1 & x_0 & \dots & x_0^n \\ 1 & x_1 & \dots & x_1^n \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \dots & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

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 (3.1) imposent la condition

$$L_i(x_j) = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{sinon} \end{cases} \qquad \text{pour } 0 \le i, j \le n,$$

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 $i \neq j$, $0 \le i \le n$,
- $L_i(x_i) = 1$.

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 satisfaite pour tout $x \in \mathbb{R}$ alors en particulier elle doit être satisfaite 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 α_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 (3.1) et ceci pour n'importe quelles valeurs $y_0, y_1, y_2, \dots, y_n$ données. Ceci montre que le système linéaire (3.2) 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

$$p_n(x) = \sum_{i=0}^{n} y_i L_i(x) \qquad \text{où} \quad L_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j}.$$

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 sort 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 [1]

$$\begin{aligned} \omega_0(x) &= 1, \\ \omega_k(x) &= \prod_{i=0}^{k-1} (x - x_i) = (x - x_{k-1})\omega_{k-1}(x), \qquad \forall k = 1, \dots, n. \end{aligned}$$

Il est facile de vérifier que

1. Notons que le dernier point x_n n'intervient pas dans la construction de cette base.

- $\omega_k(x) \in \mathbb{R}_n[x]$,
- la famille $\{\omega_0, \omega_1, \omega_2, \ldots, \omega_n\}$ est génératrice de $\mathbb{R}_n[x]$
- la famille $\{\omega_0, \omega_1, \omega_2, \ldots, \omega_n\}$ est libre.

Par conséquent, la famille $\{\omega_0, \omega_1, \omega_2, \ldots, \omega_n\}$ forme une base de $\mathbb{R}_n[x]$.
Si on choisit comme base de $\mathbb{R}_n[x]$ la famille $\{\omega_0, \omega_1, \omega_2, \ldots, \omega_n\}$, le problème du calcul du polynôme d'interpolation p_n est alors ramené au calcul des coefficients $\{\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_n\}$ tels que

$$p_n(x) = \sum_{i=0}^{n} \alpha_i \omega_i(x).$$

Si on a calculé les $n+1$ coefficients $\{\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_n\}$ et on ajoute un point d'interpolation, il n'y a plus à calculer que le coefficient α_{n+1} car la nouvelle base est déduite de l'autre base en ajoutant simplement le polynôme ω_{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 $x_0, \ldots, x_n$ est définie par récurrence en utilisant deux différences divisées d'ordre $n-1$ comme suit :

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

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	$f[x_{i-1}, x_i]$	$f[x_{i-2}, x_{i-1}, x_i]$	$f[x_{i-3}, x_{i-2}, x_{i-1}, x_i]$	$f[x_{i-4}, x_{i-3}, x_{i-2}, x_{i-1}, x_i]$	$\ldots$
0	x_0	$\boxed{y_0}$					
1	x_1	y_1	$\boxed{f[x_0, x_1]}$				
2	x_2	y_2	$f[x_1, x_2]$	$\boxed{f[x_0, x_1, x_2]}$			
3	x_3	y_3	$f[x_2, x_3]$	$f[x_1, x_2, x_3]$	$\boxed{f[x_0, x_1, x_2, x_3]}$		
4	x_4	y_4	$f[x_3, x_4]$	$f[x_2, x_3, x_4]$	$f[x_1, x_2, x_3, x_4]$	$\boxed{f[x_0, x_1, x_2, x_3, x_4]}$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\boxed{\ddots}$

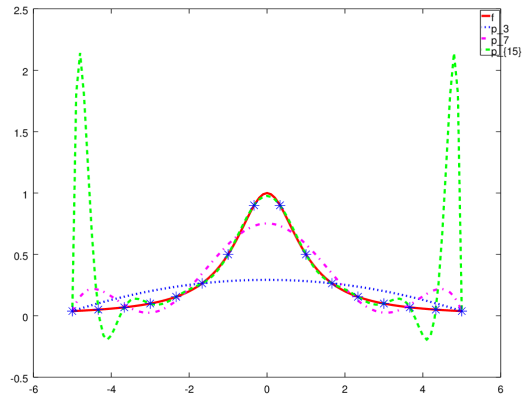
Soit $\{(x_i, y_i)\}_{i=0}^{n}$ un ensemble de $n+1$ points distincts. Le polynôme d'interpolation p_n sous la forme de NEWTON est donné par

$$p_n(x) = \sum_{i=0}^{n} \omega_i(x) f[x_0, \ldots, x_i]$$

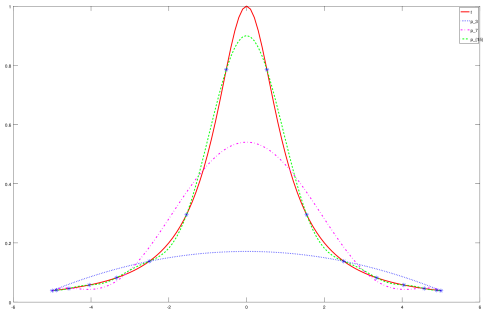
où

$$\begin{aligned}
\omega_0(x) &= 1, \\
\omega_k(x) &= \prod_{i=1}^{k-1} (x - x_i) = (x - x_{k-1})\omega_{k-1}(x), \qquad \forall k = 1, \ldots, n; \\
f[x_k] &= y_k, \qquad \forall k = 0, \ldots, n, \\
f[x_0, \ldots, x_k] &\equiv \frac{f[x_1, \ldots, x_k] - f[x_0, \ldots, x_{k-1}]}{x_k - x_0}, \qquad \forall k = 1, \ldots, n.
\end{aligned}$$

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.



(a) Distribution équirepartie des nœuds



(b) Nœuds de CHEBYSHEV-GAUSS-LOBATTO

FIGURE 3.1. – Interpolation de LAGRANGE, exemple de RUNGE

Remarque (Les défauts de l'interpolation polynomiale)
Malheureusement les polynômes d'interpolation ne forment pas une suite convergente vers une fonction analytique f.
Si $y_i = f(x_i)$ pour $i = 1, 2, \ldots, n$, $f : I \to \mathbb{R}$ étant une fonction donnée de classe $\mathscr{C}^n(I)$ où I est le plus petit intervalle contenant les nœuds 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

$$E_{n-1}(x) \overset{\text{def}}{=} f(x) - p_{n-1}(x) = \frac{f^{(n)}(\xi)}{n!}\omega_n(x)$$

avec $p_{n-1} \in \mathbb{R}_{n-1}[x]$ le polynôme d'interpolation.
Dans le cas d'une distribution uniforme de nœuds, *i.e.* quand $x_i = x_{i-1} + h$ avec $i = 1, 2, \ldots, n$ et $h > 0$ et x_0 donnés, on a

$$|\omega_n(x)| \leq (n-1)!\frac{h^n}{4}$$

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 $\max_{x \in I} |E_{n-1}(x)| \xrightarrow[n \to +\infty]{} +\infty$. 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 de l'exemple ci-dessous.

EXEMPLE
Soit la fonction $f : [-5, 5] \to \mathbb{R}$ 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 3.1a montre ses polynômes interpolants de degrés 3, 5 et 10 pour une distribution équirepartie 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 3.1b)

$$x_i = \frac{a+b}{2} - \frac{b-a}{2}\cos\left(\frac{\pi}{n-1}(i-1)\right), \qquad \text{pour } i = 0, \ldots, n$$

Pour cette distribution particulière de nœuds, il est possible de montrer que, si f est dérivable sur $[a, b]$, alors p_n converge vers f quand $n \to +\infty$ 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.
Les courbes des figures 3.1a et 3.1b peuvent été obtenues par les instructions suivantes :

```
f=@(x)[1./(1+x.^2)]; % La fonction de Runge
```



```
x = [-5:.1:5]; % Pour l'affichage on evaluera 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 = polyval(polyfit(x1,y1,3),x);

% n=8 points => p in R_7[x]
x2 = [linspace(-5,5,8)];
y2 = f(x2);
y2interp = polyval(polyfit(x2,y2,7),x);

% n=16 points => p in R_15[x]
x3 = [linspace(-5,5,16)];
y3 = f(x3);
y3interp = polyval(polyfit(x3,y3,15),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 = polyval(polyfit(x1,y1,3),x);

% n=8 points => p in R_7[x]
x2 = -5*cos(pi*[0:7]/7);
y2 = f(x2);
y2interp = polyval(polyfit(x2,y2,7),x);

% n=16 points => p in R_15[x]
x3 = -5*cos(pi*[0:15]/15);
y3 = f(x3);
y3interp = polyval(polyfit(x3,y3,15),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");
```

Remarque (Splines : interpolation composite)
On a mis en évidence le fait que, quand les nœuds d'interpolation sont équirépartis, on ne peut pas garantir la convergence uniforme du polynôme interpolatoire de LAGRANGE vers f. 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 $[a;b] = \cup_{0\le i\le n-1}[x_i, x_{i+1}]$ et d'utiliser l'interpolation de LAGRANGE sur chaque sous-intervalles $[x_i, x_{i+1}]$ en utilisant m nœuds équirépartis avec m petit (généralement $m = 1$ ou 3).
Ici nous allons considérer seulement le cas $m = 1$, *i.e.* des splines linéaires.
Spline linéaire : étant donné une distribution (non nécessairement uniforme) de nœuds $x_0 < x_1 < \cdots < x_n$, 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 3.1 (*Splines* linéaires)
Étant donné $n+1$ points distincts $x_0, \dots, x_n$ de $[a;b]$ avec $a = x_0 < x_1 < \cdots < x_n = b$, la fonction $\ell\colon [a;b] \to \mathbb{R}$ est une spline linéaire relative aux nœuds $\{x_i\}$ si

$$\begin{cases} \ell(x)|_{[x_i;x_i+1]} \in \mathbb{R}_1, \quad i = 1, 1, \dots, n-1, \\ \ell \in \mathscr{C}^0([a;b]). \end{cases}$$

Autrement dit, dans chaque sous-intervalle $[x_i; x_i+1]$, la fonction $\ell\colon [x_i, x_{i+1}] \to \mathbb{R}$ 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 ℓ 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 $i = 1, 1, \dots, n-1$.

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 $i = 0, \dots, 2$. On calculera enfin la spline linéaire associée aux mêmes points.
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 linéaireon 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	$\boxed{1}$		
1	0	0	$\boxed{-1}$	
2	1	1	1	$\boxed{1}$

On a alors

$$\begin{aligned} p_2(x) &= \sum_{i=0}^{2} \omega_i(x) f[x_0, \dots, x_i] \\ &= \omega_0(x) f[x_0] + \omega_1(x) f[x_0, x_1] + \omega_2(x) f[x_0, x_1, x_2] \\ &= \omega_0(x) - \omega_1(x) + \omega_2(x) \\ &= 1 - (x+1) + x(x+1) = x^2. \end{aligned}$$

Spline linéaire.

$$s_1(x) = \begin{cases} -x & \text{si } -1 \le x \le 0, \\ x & \text{si } 0 \le x \le 1. \end{cases}$$

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 $i = 0, \dots, 2$. Calculer ensuite la spline linéaire associée aux même points.

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[2] 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

$$p_2(x) = y_0 L_0(x) + y_1 L_1(x) + y_2 L_2(x) = \frac{x(x-\pi)}{\frac{\pi}{2}\left(\frac{\pi}{2} - \pi\right)} = -\frac{4}{\pi^2} x(x-\pi).$$

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	$\boxed{0}$		
1	$\frac{\pi}{2}$	1	$\boxed{\frac{2}{\pi}}$	
2	π	0	$-\frac{2}{\pi}$	$\boxed{-\frac{4}{\pi^2}}$

On a alors

$$\begin{aligned} p_2(x) &= \sum_{i=0}^{2} \omega_i(x) f[x_0, \dots, x_i] \\ &= \omega_0(x) f[x_0] + \omega_1(x) f[x_0, x_1] + \omega_2(x) f[x_0, x_1, x_2] \\ &= \frac{2}{\pi} \omega_1(x) - \frac{4}{\pi^2} \omega_2(x) \\ &= \frac{2}{\pi} x - \frac{4}{\pi^2} x\left(x - \frac{\pi}{2}\right) \\ &= -\frac{4}{\pi^2} x(x-\pi). \end{aligned}$$

Spline linéaire.

$$s_1(x) = \begin{cases} \frac{2}{\pi} x & \text{si } 0 \le x \le \frac{\pi}{2}, \\ \frac{2}{\pi}(x - \pi) & \text{si } \frac{\pi}{2} \le x \le \pi. \end{cases}$$

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

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & \frac{\pi}{2} & \frac{\pi^2}{4} & \frac{\pi^3}{8} \\ 1 & \pi & \pi^2 & \pi^3 \\ 1 & \frac{3\pi}{2} & \frac{9\pi^2}{4} & \frac{27\pi^3}{8} \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ -1 \end{pmatrix}$$

2. Par la méthode du pivot de Gauss on obtient

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ 1 & \frac{\pi}{2} & \frac{\pi^2}{4} & 1 \\ 1 & \pi & \pi^2 & 0 \end{array}\right) \xrightarrow[L_3 \leftarrow L_3 - L_1]{L_2 \leftarrow L_2 - L_1} \left(\begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ 0 & \frac{\pi}{2} & \frac{\pi^2}{4} & 1 \\ 0 & \pi & \pi^2 & 0 \end{array}\right) \xrightarrow{L_3 \leftarrow L_3 - 2L_2} \left(\begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ 0 & \frac{\pi}{2} & \frac{\pi^2}{4} & 1 \\ 0 & 0 & \frac{\pi^2}{2} & -2 \end{array}\right)$$

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

$$p_3(x) = y_0L_0(x) + y_1L_1(x) + y_2L_2(x) + y_3L_3(x) = \frac{x(x-\pi)\left(x-\frac{3\pi}{2}\right)}{\frac{\pi}{2}\left(\frac{\pi}{2}-\pi\right)\left(\frac{\pi}{2}-\frac{3\pi}{2}\right)} - \frac{x\left(x-\frac{\pi}{2}\right)(x-\pi)}{\frac{3\pi}{2}\left(\frac{3\pi}{2}-\frac{\pi}{2}\right)\left(\frac{3\pi}{2}-\pi\right)}$$
$$= \frac{4}{\pi^3}x(x-\pi)\left(x-\frac{3\pi}{2}\right) - \frac{4}{3\pi^3}x\left(x-\frac{\pi}{2}\right)(x-\pi).$$

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	$\boxed{0}$			
1	$\frac{\pi}{2}$	1	$\boxed{\frac{2}{\pi}}$		
2	π	0	$-\frac{2}{\pi}$	$\boxed{-\frac{4}{\pi^2}}$	
3	$\frac{3\pi}{2}$	-1	$-\frac{2}{\pi}$	0	$\boxed{\frac{8}{3\pi^3}}$

On a alors

$$\begin{aligned} p_3(x) &= \sum_{i=0}^{3} \omega_i(x) f[x_0,\dots,x_i] \\ &= p_2(x) + \omega_3(x) f[x_0,x_1,x_2,x_3] \\ &= -\frac{4}{\pi^2}x(x-\pi) + \frac{8}{3\pi^3}\omega_3(x) \\ &= -\frac{4}{\pi^2}x(x-\pi) + \frac{8}{3\pi^3}x\left(x-\frac{\pi}{2}\right)(x-\pi) \\ &= \frac{8}{3\pi^3}x(x^2-3\pi x+2\pi^2). \end{aligned}$$

Spline linéaire.

$$s_1(x) = \begin{cases} \frac{2}{\pi}x & \text{si } 0 \le x \le \frac{\pi}{2}, \\ -\frac{2}{\pi}(x-\pi) & \text{si } \frac{\pi}{2} \le x \le \pi \\ -\frac{2}{\pi}(x-\pi) & \text{si } \pi \le x \le \frac{3\pi}{2}. \end{cases}$$

Remarque

Si n est petit il est souvent plus simple de calculer directement les coefficients a_0, a_1, …, a_n en résolvant le système linéaire (3.2).

3.2. Interpolation non polynomiale

Une généralisation de l'interpolation polynomiale consiste à chercher la fonction interpolant les $n+1$ points donnés non pas dans $\mathbb{R}_n[x]$ mais dans un autre espace vectoriel $\mathcal{V}$ engendré par $n+1$ fonctions libres $\{\varphi_j,\ j=0,\dots,n\}$. On peut considérer par exemple des fonctions trigonométriques $\varphi_j(x) = \cos(jx)$, des fonctions exponentielles $\varphi_j(x) = e^{jx}$ etc. Le choix des fonctions $\{\varphi_j\}$ est en pratique dicté par la forme supposée de la loi décrivant les données.
On considère donc un ensemble de $(n+1)$ points $\{(x_i,y_i)\}_{i=0}^n$ et on cherche une fonction $f(x) = \sum_{j=0}^n a_j\varphi_j(x)$ telle que $f(x_i) = y_i$ où $a_0, a_1, a_2, \dots, a_n$ sont des coefficients qui devront être déterminés. Les $(n+1)$ relations s'écrivent alors

$$\begin{cases} a_0\varphi_0(x_0) + a_1\varphi_1(x_0) + \dots a_n\varphi_n(x_0) = y_0 \\ a_0\varphi_0(x_1) + a_1\varphi_1(x_1) + \dots a_n\varphi_n(x_1) = y_1 \\ \dots \\ a_0\varphi_0(x_n) + a_1\varphi_1(x_n) + \dots a_n\varphi_n(x_n) = y_n \end{cases}$$

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

$$\begin{pmatrix} \varphi_0(x_0) & \varphi_1(x_0) & \dots & \varphi_n(x_0) \\ \varphi_0(x_1) & \varphi_1(x_1) & \dots & \varphi_n(x_1) \\ \vdots & \vdots & & \vdots \\ \varphi_0(x_n) & \varphi_1(x_n) & \dots & \varphi_n(x_n) \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}. \tag{3.3}$$

Ainsi, le problème consistant à chercher la fonction f peut se réduire à résoudre le système linéaire (3.3).

Étant donné $n+1$ points distincts $x_0, \dots, x_n$ et $n+1$ valeurs correspondantes $y_0, \dots, y_n$, il existe une unique fonction f de l'espace vectoriel $\mathcal{V}$ de base $\{\varphi_0(x), \varphi_1(x), \dots, \varphi_n(x)\}$ telle que $f(x_i) = y_i$, pour $i = 0, \dots n$ qu'on peut écrire sous la forme

$$f(x) = \sum_{i=0}^{n} a_i \varphi_j(x) \qquad \text{avec} \qquad \begin{pmatrix} \varphi_0(x_0) & \varphi_1(x_0) & \dots & \varphi_n(x_0) \\ \varphi_0(x_1) & \varphi_1(x_1) & \dots & \varphi_n(x_1) \\ \vdots & \vdots & & \vdots \\ \varphi_0(x_n) & \varphi_1(x_n) & \dots & \varphi_n(x_n) \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

Deux cas particuliers :

- ⋆ si $\varphi_j(x) = x^j$ on retrouve le cas du fitting polynomial,
- ⋆ si $\varphi_j(x) = \cos(x)$ ou $\phi_j(x) = \sin(x)$ on parle d'interpolation trigonométrique et il n'est pas nécessaire de calculer les coefficients en résolvant le système linéaire (3.3).

3.2.1. Interpolation Trigonométrique

On veut approcher une fonction périodique $f\colon [0;2\pi] \to \mathbb{C}$, *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 2π 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 :

$$\begin{cases} \cos(kx) = \frac{e^{ikx}+e^{-ikx}}{2}, \\ \sin(kx) = \frac{e^{ikx}-e^{-ikx}}{2i} = -i\frac{e^{ikx}-e^{-ikx}}{2}, \end{cases} \qquad \text{et} \qquad \begin{cases} e^{ikx} = \cos(kx) + i\sin(kx), \\ e^{-ikx} = \cos(kx) - i\sin(kx). \end{cases}$$

Ainsi

$$\begin{aligned} \tilde{f}(x) &= a_0 + \sum_{k=1}^{K} a_k \frac{e^{ikx}+e^{-ikx}}{2} - ib_k \frac{e^{ikx}-e^{-ikx}}{2} \\ &= a_0 + \sum_{k=1}^{K} \underbrace{\frac{a_k - ib_k}{2}}_{c_k} e^{ikx} + \underbrace{\frac{a_k + ib_k}{2}}_{c_{-k}} e^{i(-k)x} = \sum_{k=-K}^{K} c_k e^{ikx} \end{aligned}$$

les inconnues sont maintenant les $2K+1$ coefficients $c_k \in \mathbb{C}$ et l'on a les relations

$$\begin{cases} c_0 = a_0, \\ c_k = \frac{a_k - ib_k}{2}, & k = 1, \dots, K \\ c_{-k} = \overline{c_k} = \frac{a_k + ib_k}{2}, & k = 1, \dots, K \end{cases} \iff \begin{cases} a_0 = c_0, \\ a_k = c_k + c_{-k}, & k = 1, \dots, K \\ b_k = i(c_k - c_{-k}), & k = 1, \dots, K. \end{cases}$$

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

$$\tilde{f}(x) = \sum_{k=-(M+\mu)}^{M} c_k e^{ikx}, \qquad \mu = \begin{cases} 0 & \text{si } n \text{ est pair,} \\ 1 & \text{si } n \text{ est impair,} \end{cases}$$

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 :

$$\sum_{j=0}^{n} \left(f(x_j) e^{-imjh} \right) = \sum_{j=0}^{n} \left(\sum_{k=-(M+\mu)}^{M} c_k e^{i(k-m)jh} \right).$$

En échangeant l'ordre de sommation on obtient

$$\sum_{j=0}^{n} \left(f(x_j) e^{-imjh} \right) = \sum_{k=-(M+\mu)}^{M} \left(c_k \left(\sum_{j=0}^{n} e^{i(k-m)jh} \right) \right).$$

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

$$\sum_{j=0}^{n} \left(f(x_j) e^{-imjh} \right) = (n+1) \sum_{k=-(M+\mu)}^{M} \delta_{km} c_k$$

i.e. seul le terme $k = m$ est à prendre en considération

$$\sum_{j=0}^{n} \left(f(x_j) e^{-imjh} \right) = (n+1) c_m \qquad m = -M-\mu, \dots M.$$

Soit $\left\{ (x_j = jh, f(x_j)) \right\}_{j=0}^{n}$ un ensemble de $n+1$ points avec $h = 2\pi/(n+1)$ et $f\colon [0;2\pi] \to \mathbb{C}$ une fonction périodique. Le polynôme trigonométrique d'interpolation $\tilde{f}$ est donné par

$$\tilde{f}(x) = \sum_{k=-(M+\mu)}^{M} c_k e^{-ikx}, \qquad (M, \mu) = \begin{cases} (n/2, 0) & \text{si } n \text{ est pair,} \\ ((n-1)/2, 1) & \text{si } n \text{ est impair,} \end{cases}$$

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

$$\tilde{f}(x) = a_0 + \sum_{k=1}^{M+\mu} a_k\cos(kx) + b_k\sin(kx), \qquad (M,\mu) = \begin{cases} (n/2,0) & \text{si } n \text{ est pair,} \\ ((n-1)/2,1) & \text{si } n \text{ est impair,} \end{cases}$$

avec

$$\begin{cases} a_0 = \frac{1}{n+1}\sum_{j=0}^{n} f(x_j) & \\ a_k = \frac{2}{n+1}\sum_{j=0}^{n} f(x_j)\cos(kx_j), & k=1,\dots M+\mu, \\ b_k = \frac{2}{n+1}\sum_{j=0}^{n} f(x_j)\sin(kx_j), & k=1,\dots M+\mu, \\ a_{M+\mu} = ib_{M+\mu} & \text{si } \mu = 1. \end{cases}$$

Il est intéressant de noter que l'expression de c_k est une approximation de l'intégrale $\frac{1}{2\pi}\int_0^{2\pi} f(x)e^{-ikx}\,\mathrm{d}x$ par la méthode des rectangles à gauche composite. De la même manière, les coefficients a_k et b_k sont des approximations des intégrales $\frac{1}{\pi}\int_0^{2\pi} f(x)\cos(kx)\,\mathrm{d}x$ et $\frac{1}{\pi}\int_0^{2\pi} f(x)\sin(kx)\,\mathrm{d}x$ 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 $f\colon [0;2\pi] \to \mathbb{R}$ 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 $\{(0,f(0)),(\pi,f(\pi))\} = \{(0,0),(\pi,-\pi^2e^{-\pi})\}$.

  Méthode directe On cherche a_0 et a_1 tels que $\tilde{f}(x) = a_0 + a_1\cos(x)$ vérifie $\tilde{f}(0) = 0$ et $\tilde{f}(\pi) = -\pi^2e^{-\pi}$:

  $$\begin{pmatrix} 1 & \cos(0) \\ 1 & \cos(\pi) \end{pmatrix}\begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 0 \\ -\pi^2e^{-\pi} \end{pmatrix} \quad i.e. \quad \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 0 \\ -\pi^2e^{-\pi} \end{pmatrix}$$

  ainsi $a_0 = -\frac{\pi^2e^{-\pi}}{2} = -a_1$ et

  $$\tilde{f}(x) = \frac{\pi^2e^{-\pi}}{2}(-1+\cos(x)).$$

  Méthode "Fourier" n étant impair, $M = (n-1)/2 = 0$ et $\mu = 1$ et

  $$\tilde{f}(x) = \sum_{k=-1}^{0} c_k e^{ikx6} = c_{-1}e^{-ix} + c_0$$

  On doit alors calculer les deux coefficients de FOURIER c_{-1} et c_0 :

  $$c_{-1} = \frac{1}{n+1}\sum_{j=0}^{1} f(x_j)e^{-ix_j} = \frac{1}{2}\left(f(x_0)e^{-ix_0} + f(x_1)e^{-ix_1}\right) = \frac{1}{2}\left(-\pi^2e^{-\pi}e^{-i\pi}\right) = \frac{\pi^2e^{-\pi}}{2}$$
  $$c_0 = \frac{1}{n+1}\sum_{j=0}^{1} f(x_j) = \frac{1}{2}\left(f(x_0)+f(x_1)\right) = \frac{1}{2}\left(-\pi^2e^{-\pi}\right) = -\frac{\pi^2e^{-\pi}}{2}$$

  ainsi

  $$\tilde{f}(x) = \frac{\pi^2e^{-\pi}}{2}(e^{-ix}-1).$$

  Lien entre les deux solutions $b_1 = 0$ (il n'y a pas de $\sin(x)$ dans la base choisie) donc $c_0 = a_0$, $c_1 = \frac{a_1 - ib_1}{2} = \frac{a_1}{2}$, $c_{-1} = \frac{a_1-ib_1}{2} = \frac{a_1}{2}$ et $\cos(x) = \frac{e^{ix}+e^{-ix}}{2}$ ainsi on a bien

  $$\tilde{f}(x) = \frac{\pi^2e^{-\pi}}{2}(-1+\cos(x)) = \frac{\pi^2e^{-\pi}}{2}(e^{-ix}-1).$$

- 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 $\{(0,f(0)),(\frac{2\pi}{3},f(\frac{2\pi}{3})),(\frac{4\pi}{3},f(\frac{4\pi}{3}))\} = \left\{(0,0),(\frac{2\pi}{3},-\frac{8\pi^2}{9}e^{-\frac{2\pi}{3}}),(\frac{4\pi}{3},-\frac{8\pi^2}{9}e^{-\frac{4\pi}{3}})\right\}$.

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 :

$$\begin{aligned}
c_{-1} &= \frac{1}{n+1}\sum_{j=0}^{2} f(x_j)e^{-ix_j} = \frac{1}{3}\left(f(x_0)e^{-ix_0} + f(x_1)e^{-ix_1} + f(x_2)e^{-ix_2}\right) = \frac{1}{3}\left(-\frac{8\pi^2}{9}e^{-(i+1)\frac{2\pi}{3}} - \frac{8\pi^2}{9}e^{-(i+1)\frac{4\pi}{3}}\right) \\
&= -\frac{8\pi^2}{27}\left(e^{-(i+1)\frac{2\pi}{3}} + e^{-(i+1)\frac{4\pi}{3}}\right) = -\frac{8\pi^2}{27}e^{-(i+1)\frac{2\pi}{3}}\left(1 + e^{-(i+1)\frac{2\pi}{3}}\right) \\
c_0 &= \frac{1}{n+1}\sum_{j=0}^{2} f(x_j) = \frac{1}{3}\left(f(x_0) + f(x_1) + f(x_2)\right) = \frac{1}{3}\left(-\frac{8\pi^2}{9}e^{-\frac{2\pi}{3}} - \frac{8\pi^2}{9}e^{-\frac{4\pi}{3}}\right) = -\frac{8\pi^2}{27}e^{-\frac{2\pi}{3}}\left(1 + e^{-\frac{2\pi}{3}}\right) \\
c_1 &= \frac{1}{n+1}\sum_{j=0}^{2} f(x_j)e^{ix_j} = \frac{1}{3}\left(f(x_0)e^{ix_0} + f(x_1)e^{ix_1} + f(x_2)e^{ix_2}\right) = \frac{1}{3}\left(-\frac{8\pi^2}{9}e^{(i-1)\frac{2\pi}{3}} - \frac{8\pi^2}{9}e^{(i-1)\frac{4\pi}{3}}\right) \\
&= -\frac{8\pi^2}{27}\left(e^{(i-1)\frac{2\pi}{3}} + e^{(i-1)\frac{4\pi}{3}}\right) = -\frac{8\pi^2}{27}e^{(i-1)\frac{2\pi}{3}}\left(1 + e^{(i-1)\frac{2\pi}{3}}\right)
\end{aligned}$$

ainsi

$$\begin{aligned}
\tilde{f}(x) &= -\frac{8\pi^2}{27}e^{-(i+1)\frac{2\pi}{3}}\left(1 + e^{-(i+1)\frac{2\pi}{3}}\right)e^{-ix} - \frac{8\pi^2}{27}e^{-\frac{2\pi}{3}}\left(1 + e^{-\frac{2\pi}{3}}\right) - \frac{8\pi^2}{27}e^{(i-1)\frac{2\pi}{3}}\left(1 + e^{(i-1)\frac{2\pi}{3}}\right)e^{ix} \\
&= -\frac{8\pi^2}{27}e^{-\frac{2\pi}{3}}\left[e^{-i\frac{2\pi}{3}}\left(1 + e^{-(i+1)\frac{2\pi}{3}}\right)e^{-ix} + \left(1 + e^{-\frac{2\pi}{3}}\right) + e^{i\frac{2\pi}{3}}\left(1 + e^{(i-1)\frac{2\pi}{3}}\right)e^{ix}\right] \\
&= -\frac{8\pi^2}{27}e^{-\frac{2\pi}{3}}\left[e^{-i(\frac{2\pi}{3}+x)}\left(1 + e^{-(i+1)\frac{2\pi}{3}}\right) + \left(1 + e^{-\frac{2\pi}{3}}\right) + e^{i(\frac{2\pi}{3}+x)}\left(1 + e^{(i-1)\frac{2\pi}{3}}\right)\right].
\end{aligned}$$

3.3. Exercices

Interpolation polynomiale

☆ Exercice 3.1
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)
   [l,c]=size(P);
   V = ones(l,l);
   V(:,2:l) = P(:,1).^(1:l-1);
   alpha = V\P(:,2);
   y=zeros(size(x));
   for i=1:l
      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:l)= P(:,1).^(1:l-1)` correspond à la boucle

```
for j=1:l
   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

$$p(x) = \sum_{i=1}^{n} \alpha_i x^{i-1} = \alpha_1 + x\Big(\alpha_2 + x\Big(\alpha_3 + \cdots + x(\alpha_{n-1} + \alpha_n x)\ldots\Big)\Big).$$

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 :

$$\begin{cases} b_n = \alpha_n \\ b_k = \alpha_k + x b_{k+1} \quad \text{pou } k = n-1,\ldots,1 \end{cases}$$

et $b_0 = p(x)$. On modifie alors notre fonction comme suit

```
function [y]=naive(P,x)
   [l,c]=size(P);
   V = ones(l,l);
   V(:,2:l) = P(:,1).^(1:l-1);
   alpha = V\P(:,2);
   y=zeros(size(x));
   for k=l:-1:1
      y=alpha(k)+x.*y;
   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]=naivePoly(P)
   [l,c]=size(P);
   V = ones(l,l);
   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:l,j)=(A(j:l,j-1)-A(j-1:l-1,j-1))./(P(j:l,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:l,j)=(A(j:l,j-1)-A(j-1:l-1,j-1))./(P(j:l,1)-P(1:l-j+1,1));` correspond à la boucle

```
for i=j:l
   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

$$\begin{aligned} p(x) &= \sum_{i=1}^{n} \beta_i \omega_i(x) = \beta_1 + \beta_2(x-x_1) + \beta_3(x-x_1)(x-x_2) + \cdots + \beta_n(x-x_1)\cdots(x-x_{n-1}) \\ &= \beta_1 + (x-x_1)\Big(\beta_2 + (x-x_2)\Big(\beta_3 + \cdots + (x-x_{n-2})(\beta_{n-1} + \beta_n(x-x_{n-1}))\ldots\Big)\Big). \end{aligned}$$

Celui-ci permet d'évaluer de manière efficace un polynôme en un point en utilisant l'algorithme de division synthétique suivant :

$$\begin{cases} b_n = (x-x_{n-1})\beta_n \\ b_k = \beta_k + (x-x_k)b_{k+1} \quad \text{pou } k = n-1,\ldots,1 \end{cases}$$

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(l);
   A(:,1)=P(:,2);
   for j=2:l
      A(j:l,j)=(A(j:l,j-1)-A(j-1:l-1,j-1))./(P(j:l,1)-P(1:l-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 3.2
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 due 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 $\mathscr{C}=\{1,x,x^2,x^3\}$ la base canonique de $\mathbb{R}_3[x]$ et on cherche $(a_0,a_1,a_2,a_3)=\text{coord}(P,\mathscr{C})$, *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 :
  $$\left(\begin{array}{cccc|c} 1&0&0&0&2\\ 1&1&1&1&1\\ 1&2&4&8&2\\ 1&3&9&27&3 \end{array}\right) \xrightarrow[L_4\leftarrow L_4-L_1]{\substack{L_2\leftarrow L_2-L_1\\ L_3\leftarrow L_3-L_1}} \left(\begin{array}{cccc|c} 1&0&0&0&2\\ 0&1&1&1&-1\\ 0&2&4&8&0\\ 0&3&9&27&1 \end{array}\right) \xrightarrow[L_4\leftarrow L_4-3L_2]{\substack{L_1\leftarrow L_1-0L_2\\ L_3\leftarrow L_3-2L_2}} \left(\begin{array}{cccc|c} 1&0&0&0&2\\ 0&1&1&1&-1\\ 0&0&2&6&2\\ 0&0&6&24&4 \end{array}\right)$$
  $$\xrightarrow[L_4\leftarrow L_4-3L_3]{\substack{L_1\leftarrow L_1-0L_3\\ L_2\leftarrow L_2-L_3/2}} \left(\begin{array}{cccc|c} 1&0&0&0&2\\ 0&1&0&-2&-2\\ 0&0&2&6&2\\ 0&0&0&6&-2 \end{array}\right) \xrightarrow[L_3\leftarrow L_3-L_4]{\substack{L_1\leftarrow L_1-0L_4\\ L_2\leftarrow L_2+L_4/3}} \left(\begin{array}{cccc|c} 1&0&0&0&2\\ 0&1&0&0&-8/3\\ 0&0&2&0&4\\ 0&0&0&6&-2 \end{array}\right)$$
  donc $a_3=-\frac{1}{3}$, $a_2=2$, $a_1=-\frac{8}{3}$ et $a_0=2$ et on trouve $P(x)=2-\frac{8}{3}x+2x^2-\frac{1}{3}x^3$:

```
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 :

$$\left(\begin{array}{ccc|c} 1 & 1 & 1 & -1 \\ 2 & 4 & 8 & 0 \\ 3 & 9 & 27 & 1 \end{array}\right) \xrightarrow[L_3\leftarrow L_3-3L_1]{L_2\leftarrow L_2-2L_1} \left(\begin{array}{ccc|c} 1 & 1 & 1 & -1 \\ 0 & 2 & 6 & 2 \\ 0 & 6 & 24 & 4 \end{array}\right) \xrightarrow[L_2\leftarrow L_2/2]{\substack{L_1\leftarrow L_1-L_2\\ L_3\leftarrow L_3-6L_2}} \left(\begin{array}{ccc|c} 1 & 0 & -2 & -2 \\ 0 & 1 & 3 & 1 \\ 0 & 0 & 6 & -2 \end{array}\right) \xrightarrow[L_3\leftarrow L_3/6]{\substack{L_1\leftarrow L_1+2L_3\\ L_2\leftarrow L_2-3L_3}} \left(\begin{array}{ccc|c} 1 & 0 & 0 & -8/3 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & -1/3 \end{array}\right)$$

donc $a_3 = -\frac{1}{3}$, $a_2 = 2$, $a_1 = -\frac{8}{3}$ et $a_0 = 2$ et on trouve $P(x) = 2 - \frac{8}{3}x + 2x^2 - \frac{1}{3}x^3$.

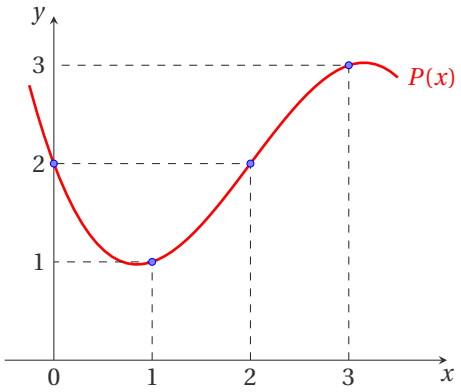
- **Méthode de Lagrange**

  On considère $\mathscr{L} = \{L_0, L_1, L_2, L_3\}$ une base de $\mathbb{R}_3[x]$ telle que $\text{coord}(P, \mathscr{L}) = (y_0, y_1, y_2, y_3)$, *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 $\mathscr{N} = \{\omega_0, \omega_1, \omega_2, \omega_3\}$ une base de $\mathbb{R}_3[x]$ telle que $\text{coord}(P, \mathscr{N}) = (y_0, f[x_0, x_1], f[x_0, x_1, x_2], f[x_0, x_1, x_2, x_3])$, *i.e.* $P(x) = \sum_{i=0}^{3} f[x_0, \ldots, x_i]\omega_i(x)$.

  La base de Newton est définie récursivement comme suit :

$$\omega_0(x) = 1; \qquad \text{pour } k = 1, \ldots, n \quad \omega_k(x) = \omega_{k-1}(x)(x - x_{k-1}).$$

  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	$\boxed{2}$			
1	1	1	$\boxed{-1}$		
2	2	2	1	$\boxed{1}$	
3	3	3	1	0	$\boxed{-\frac{1}{3}}$

  On a alors

$$\begin{aligned} P_3(x) &= \sum_{i=1}^{3} f[x_0, \ldots, x_i]\omega_i(x) \\ &= y_0\omega_0(x) + f[x_0, x_1]\omega_1(x) + f[x_0, x_1, x_2]\omega_2(x) + f[x_0, x_1, x_2, x_3]\omega_3(x) \\ &= 2\omega_0(x) - \omega_1(x) + \omega_2(x) - \frac{1}{3}\omega_3(x) \\ &= 2 - x + x(x-1) - \frac{1}{3}x(x-1)(x-2) \\ &= -\frac{1}{3}x^3 + 2x^2 - \frac{8}{3}x + 2. \end{aligned}$$

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	$\boxed{1}$			
1	2	2	$\boxed{1}$		
2	3	3	1	$\boxed{0}$	
3	0	2	$-\frac{1}{3}$	$-\frac{1}{3}$	$\boxed{-\frac{1}{3}}$

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.

Partage d'un message secret : *comment envoyer un message secret avec plusieurs espions sans pour autant que ceux-ci ne connaissent le contenu du message envoyé?*

Imaginons que l'on désire envoyer un message secret. Par codage, on peut remplacer ce message par un nombre, appelons-le n.

Considérons un polynôme $P(X) = a_k X^k + \cdots + a_1 X + n$ de degré k dont le terme indépendant vaut exactement n, autrement dit $P(0) = n$. Un corollaire du théorème fondamental de l'algèbre stipule que le polynôme P est complètement caractérisé par les valeurs qu'il prend en $k+1$ points, par exemple en $X = 1,2,\dots,k+1$.

On engage alors au moins $k+1$ espions (mieux en engager un peu plus au cas où certains seraient capturés par les «ennemis»). On donne au i-ème espion le nombre $P(i)$. Les espions se dispersent (par exemple, pour passer les lignes ennemies). Une fois qu'au moins $k+1$ espions sont arrivés à destination, il est aisé de reconstituer le polynôme et ainsi retrouver la valeur secrète n (on a *un système d'au moins $k+1$ équations linéaires* pour retrouver les $k+1$ coefficients de P).

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/`

Exercice 3.3

Calculer le message secret n si $k = 2$ et on envoie 4 espions avec les messages suivants :

Espion	1	2	3	4
Message	45	50	57	66

Correction

On doit interpoler l'ensemble $\{(1,45),(2,50),(3,57),(4,66)\}$ constitué de 4 points. Cependant on nous dit de chercher un polynôme de degré au plus 2 (et non pas 3 comme on pourrait s'y attendre). Cela signifie qu'ils ont envoyé un espion de plus en cas de "perte". Donc, pour nos calculs, on utilisera seulement 3 points parmi les 4 données et on vérifiera à posteriori que le polynôme obtenu interpole aussi le point négligé.

On choisit par exemple d'interpoler l'ensemble $\{(1,45),(2,50),(3,57)\}$: on cherche un polynôme de degré au plus 2 tel que $P(1) = 45$, $P(2) = 50$, $P(3) = 57$ et $P(4) = 66$. Construire P signifie trouver ses coordonnées dans une base de $\mathbb{R}_2[x]$. On considère trois méthodes qui sont basées sur trois choix différents de bases de $\mathbb{R}_2[x]$:

- **Méthode directe (naïve)**
  On considère $\mathscr{C} = \{1, x, x^2\}$ la base canonique de $\mathbb{R}_2[x]$ et on cherche $(a,b,c) = \text{coord}(P,\mathscr{C})$, *i.e.* a,b,c tels que $P(x) = a + bx + cx^2$. (le message secret est $P(0) = a$).

Il s'agit de trouver les 3 coefficients a, b, c solution du système linéaire

$$\begin{cases} P(1)=45, \\ P(2)=50, \\ P(3)=57, \\ P(4)=66 \end{cases} \quad i.e. \quad \begin{cases} a+b+c=45 \\ a+2b+2^2c=50 \\ a+3b+3^2c=57 \\ a+4b+4^2c=66 \end{cases}$$

Puisqu'on a envoyé un espion de trop, on a 4 équations et 3 inconnues : le système est sur-déterminé.
Négligeons pour le moment la dernière équation et résolvons avec Gauss

$$\begin{cases} a+b+c=45 \\ a+2b+4c=50 \\ a+3b+9c=57 \end{cases} \xrightarrow[\substack{L_3\leftarrow L_3-L_1 \\ \text{Étape } j=1}]{L_2\leftarrow L_2-L_1} \begin{cases} a+b+c=45 \\ b+3c=5 \\ 2b+8c=12 \end{cases} \xrightarrow[\text{Étape } j=2]{L_3\leftarrow L_3-2L_2} \begin{cases} a+b+c=45 \\ b+3c=5 \\ 2c=2 \end{cases} \implies \begin{array}{c} c=1 \\ b=5-3c=2 \\ a=45-b-c=42 \end{array}$$

Vérifions si la dernière équation est bien satisfaite :

$$42+2\times 4+1\times 4^2=66.$$

Le polynôme recherché est ainsi $P(x)=42+2x+x^2$ et le message secret est donc $P(0)=a=42$.

- **Méthode de Lagrange** Puisqu'on a envoyé un espion de trop, négligeons pour l'instant le point $(4,66)$. On construit la base de Lagrange de $\mathbb{R}_2[x]$ telle que $(45,50,57)=\text{coord}(P,\mathscr{L})$, *i.e.* $45,50,57$ tels que $P(x)=45L_0(x)+50L_1(x)+57L_2(x)$. On a

$$\begin{aligned} P(x) &= y_0L_0(x)+y_1L_1(x)+y_2L_2(x) \\ &= 45\frac{(x-2)(x-3)}{(1-2)(1-3)}+50\frac{(x-1)(x-3)}{(2-1)(2-3)}+57\frac{(x-1)(x-2)}{(3-1)(3-2)} \\ &= \frac{45}{2}(x-2)(x-3)-50(x-1)(x-3)+\frac{57}{2}(x-1)(x-2). \end{aligned}$$

Bien sur $P(4)=66$ (un espion était redondant) et le message secret est donc

$$P(0)=\frac{45}{2}\times 6-50\times 3+\frac{57}{2}\times 2=135-150+57=42.$$

Méthode de Newton. Puisqu'on a envoyé un espion de trop, négligeons pour l'instant le point $(4,66)$.
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	$\boxed{45}$		
1	2	50	$\boxed{5}$	
2	3	57	7	$\boxed{1}$

On a alors

$$\begin{aligned} P(x) &= \sum_{i=1}^{2}\omega_i(x)f[x_0,\dots,x_i] \\ &= \omega_0(x)f[x_0]+\omega_1(x)f[x_0,x_1]+\omega_2(x)f[x_0,x_1,x_2] \\ &= 45\omega_0(x)+5\omega_1(x)+\omega_2(x) \\ &= 45+5(x-1)+(x-1)(x-2). \end{aligned}$$

Bien sur $P(4)=45+5\times 3+3\times 2=66$ (un espion était redondant) et le message secret est donc

$$P(0)=45-5+2=42.$$

Exercice 3.4

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

   $$p_2(x) = y_0 L_0(x) + y_1 L_1(x) + y_2 L_2(x) = 1\frac{\left(x-\frac{\pi}{2}\right)(x-\pi)}{\left(0-\frac{\pi}{2}\right)(0-\pi)} - 1\frac{(x-0)\left(x-\frac{\pi}{2}\right)}{(\pi-0)\left(\pi-\frac{\pi}{2}\right)} = 1 - \frac{2}{\pi}x.$$

   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	$\boxed{1}$		
1	$\frac{\pi}{2}$	0	$\boxed{-\frac{2}{\pi}}$	
2	π	-1	$-\frac{2}{\pi}$	$\boxed{0}$

   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

$$\begin{aligned} p_3(x) &= y_0L_0(x)+y_1L_1(x)+y_2L_2(x)+y_3L_3(x) = 1\frac{\left(x-\frac{\pi}{2}\right)(x-\pi)\left(x-\frac{3\pi}{2}\right)}{\left(0-\frac{\pi}{2}\right)(0-\pi)\left(0-\frac{3\pi}{2}\right)} - 1\frac{(x-0)\left(x-\frac{\pi}{2}\right)\left(x-\frac{3\pi}{2}\right)}{(\pi-0)\left(\pi-\frac{\pi}{2}\right)\left(\pi-\frac{3\pi}{2}\right)} \\ &= \frac{4}{3\pi^3}\left(x-\frac{\pi}{2}\right)\left(x-\frac{3\pi}{2}\right)(-x+\pi+3x) = 1-\frac{2}{3\pi}x-\frac{4}{\pi^2}x^2+\frac{8}{3\pi^3}x^3. \end{aligned}$$

Méthode de Newton. Il suffit de calculer une différence divisée en plus, *i.e.* ajouter une ligne au tableau précédant :

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	$\boxed{1}$			
1	$\frac{\pi}{2}$	0	$\boxed{-\frac{2}{\pi}}$		
2	π	-1	$-\frac{2}{\pi}$	$\boxed{0}$	
3	$\frac{3\pi}{2}$	0	$\frac{2}{\pi}$	$\frac{4}{\pi^2}$	$\boxed{\frac{8}{3\pi^3}}$

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 3.5

1. Construire le polynôme P qui interpole les points $(-1,2)$, $(0,1)$, $(1,2)$ et $(2,3)$.
2. Soit Q le polynôme qui interpole les points $(-1,2)$, $(0,1)$, $(1,2)$. Montrer qu'il existe un réel λ 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

$$P(x) = y_0\frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} + y_1\frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)}$$

$$
\begin{aligned}
&+ y_2 \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} + y_3 \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} \\
&= \frac{x(x-1)(x-2)}{-3} + \frac{(x+1)(x-1)(x-2)}{2} - (x+1)x(x-2) + \frac{(x+1)x(x-1)}{2} = \\
&= -\frac{1}{3}x^3 + x^2 + \frac{1}{3}x + 1.
\end{aligned}
$$

```
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

$$
\begin{aligned}
Q(x) &= y_0 \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} + y_1 \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} + y_2 \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} \\
&= x(x-1) - (x+1)(x-1) + (x+1)x \\
&= x^2 + 1.
\end{aligned}
$$

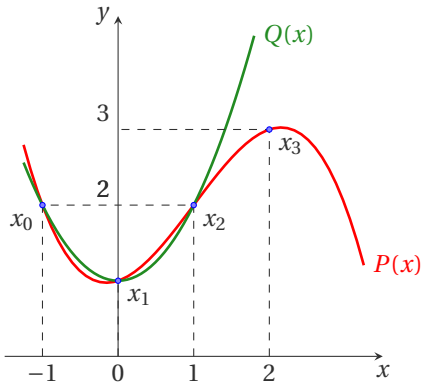
Ainsi

$$
\begin{aligned}
Q(x) - P(x) &= y_0 \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} \left[1 - \frac{x-x_3}{x_0-x_3}\right] + y_1 \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} \left[1 - \frac{x-x_3}{x_1-x_3}\right] \\
&+ y_2 \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} \left[1 - \frac{x-x_3}{x_2-x_3}\right] - y_3 \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} \\
&= -y_0 \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} - y_1 \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} \\
&- y_2 \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} - y_3 \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} \\
&= -\left[\frac{y_0}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} + \frac{y_1}{(x_1-x_0)(x_1-x_2)(x_1-x_3)}\right. \\
&\left. + \frac{y_2}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} + \frac{y_3}{(x_3-x_0)(x_3-x_1)(x_3-x_2)}\right](x-x_0)(x-x_1)(x-x_2) \\
&= \frac{(x+1)x(x-1)}{3}
\end{aligned}
$$

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 $\lambda = \frac{1}{3}$.



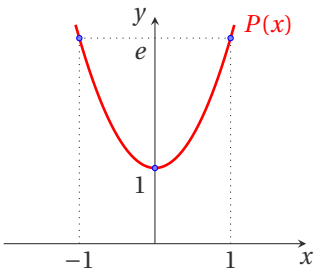
Exercice 3.6

1. Construire le polynôme P qui interpole 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 $\mathrm{Vec}\{1,x,x^2\}$ qui interpole les trois points $(-1,-1)$, $(0,0)$ et $(1,-1)$.

Correction

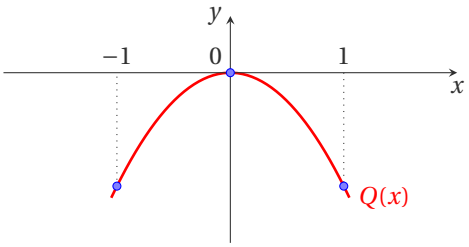
1. Dans la base de LAGRANGE le polynôme d'interpolation de degré $n=2$ s'écrit

$$\begin{aligned} P(x) &= y_0\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} + y_1\frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} + y_2\frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} = \\ &= e\frac{x(x-1)}{2} - (x+1)(x-1) + e\frac{(x+1)x}{2} = \\ &= (e-1)x^2+1. \end{aligned}$$



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 $Q\in\mathrm{Vec}\{1,x,x^2\}$ pour conclure que le polynôme p cherché est Q lui même.

Exercice 3.7

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 λ 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

$$\begin{aligned}Q(x) &= y_0\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} + y_1\frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} + y_2\frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}\\ &= \frac{x(x-1)}{2} - (x+1)(x-1) + (x+1)x\\ &= \frac{1}{2}x^2 + \frac{1}{2}x + 1.\end{aligned}$$

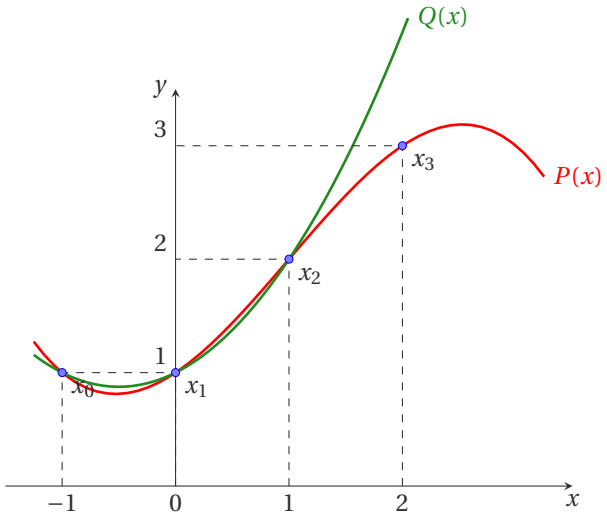
Ainsi

$$\begin{aligned}Q(x)-P(x) &= y_0\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}\left[1-\frac{x-x_3}{x_0-x_3}\right] + y_1\frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}\left[1-\frac{x-x_3}{x_1-x_3}\right]\\ &\quad + y_2\frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}\left[1-\frac{x-x_3}{x_2-x_3}\right] - y_3\frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)}\\ &= -y_0\frac{(x-x_0)(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} - y_1\frac{(x-x_0)(x-x_1)(x-x_2)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)}\\ &\quad - y_2\frac{(x-x_0)(x-x_1)(x-x_2)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} - y_3\frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)}\\ &= -\left[\frac{y_0}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} + \frac{y_1}{(x_1-x_0)(x_1-x_2)(x_1-x_3)}\right.\\ &\quad \left. + \frac{y_2}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} + \frac{y_3}{(x_3-x_0)(x_3-x_1)(x_3-x_2)}\right](x-x_0)(x-x_1)(x-x_2) = \frac{(x+1)x(x-1)}{6}\end{aligned}$$

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 3.8

1. Construire le polynôme P qui interpole les trois points $(-1,\alpha)$, $(0,\beta)$ et $(1,\alpha)$ où α et β 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

$$\begin{aligned}P(x) &= y_0\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} + y_1\frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} + y_2\frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} + \\ &= \alpha\frac{x(x-1)}{2} + \beta\frac{(x+1)(x-1)}{-1} + \alpha\frac{(x+1)x}{2} = \\ &= \frac{\alpha}{2}x(x-1) - \beta(x+1)(x-1) + \frac{\alpha}{2}x(x+1) \\ &= (\alpha-\beta)x^2+\beta.\end{aligned}$$

   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	$\boxed{\alpha}$		
1	0	β	$\boxed{(\beta-\alpha)}$	
2	1	α	$(\alpha-\beta)$	$\boxed{(\alpha-\beta)}$

   On a alors

$$\begin{aligned}p_2(x) &= \sum_{i=0}^{2}\omega_i(x)f[x_0,\dots,x_i] \\ &= \omega_0(x)f[x_0] + \omega_1(x)f[x_0,x_1] + \omega_2(x)f[x_0,x_1,x_2] \\ &= \alpha\omega_0(x) + (\beta-\alpha)\omega_1(x) + (\alpha-\beta)\omega_2(x) \\ &= \alpha + (\beta-\alpha)(x+1) + (\alpha-\beta)x(x+1) = (\alpha-\beta)x^2+\beta.\end{aligned}$$

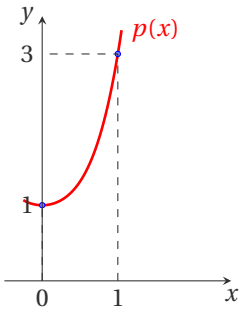
2. Si $\alpha=\beta$, $P(x)=\alpha$ qui est un polynôme de degré 0.
3. $P(-x)=P(x)$ donc P est pair. Donc P ne peut pas être de degré 1 car un polynôme de degré 1 est de la forme a_0+a_1x qui ne peut pas être pair.

Exercice 3.9
Trouver le polynôme de l'espace vectoriel $\mathrm{Vec}\{1+x^2, x^4\}$ 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 3.10
Soit $f : \mathbb{R} \to \mathbb{R}$ 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 $\{x_0 = 0, x_1 = 1, \ldots, x_n = n\}$.

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 forcement 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 forcement avec f donc $p_n \equiv f$ pour $n \geq 3$.

Exercice 3.11
Soit $f : \mathbb{R} \to \mathbb{R}$ 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 $\{0, 2, \ldots, 2i, \ldots, 18\}_{0 \leq i \leq 9}$.

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 forcement avec f donc $p_n \equiv f$ pour $n \geq 2$.

Exercice 3.12

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 $\mathscr{C}^\infty$ 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.* $p(x)=\sum_{i=0}^{n-1}f[x_0,\dots,x_i]\omega$
     La base de NEWTON est définie récursivement comme suit :

$$\omega_0(x)=0;\qquad \omega_1(x)=x-x_0;\qquad \text{pour } k=2,\dots,n\quad \omega_k(x)=\omega_{k-1}(x)(x-x_{k-1}).$$

     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 $k=0,1,2,3$:

$$\begin{cases}p(x)=a_0+a_1x+a_2x^2+a_3x^3,\\ p'(x)=a_1x+2a_2x+3a_3x^2,\\ p''(x)=2a_2+6a_3x,\\ p'''(x)=6a_3,\end{cases}\Longrightarrow\begin{cases}3=p(1)=a_0+a_1+a_2+a_3,\\ 3=p'(1)=a_1+2a_2+3a_3,\\ 3=p''(1)=2a_2+6a_3,\\ 3=p'''(1)=6a_3,\end{cases}\Longrightarrow\begin{cases}a_3=1/2,\\ a_2=0,\\ a_1=3/2,\\ a_0=1.\end{cases}$$

et ce polynôme est unique.
Soit f une fonction de classe $\mathscr{C}^\infty$ telle que $f^{(k)}(1) = 3$. Alors la fonction $g(x) \equiv f(x) - p(x)$ est de classe $\mathscr{C}^\infty$ 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 ξ 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 3.13
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

λ_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 3.14
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[3] de t. Alors $t \in [\ell; \ell+1]$ et il s'agit de définir le polynôme p interpolant les points

$$(\kappa-1, f(\kappa-1)), \qquad (\kappa, f(\kappa)), \qquad (\kappa+1, f(\kappa+1)), \qquad (\kappa+2, f(\kappa+2)),$$

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: $f : \mathbb{Z} \to \mathbb{R}$, 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**
    $L \leftarrow 1$
    for $j = 0$ to 3 **do**
      if $j \neq i$ **then**
        $L \leftarrow \dfrac{t - x_j}{x_i - x_j} \times L$
      end if
    end for
    $y \leftarrow y + f(x_i) \times L$
  end for
  return y

Exercice 3.15
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 discrétisation 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

$$f(x) = 0 \qquad \Longleftrightarrow \qquad x = p(0).$$

Utiliser cette méthode pour évaluer l'unique racine α 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

$$p_n(y) = \sum_{i=0}^{n}\left(x_i \prod_{\substack{j=0\\ j\neq i}}^{n}\frac{y-y_j}{y_i-y_j}\right).$$

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 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 $E(x+0.5)$.

Ici $n=2$ donc on a

$$p(y) = x_0\frac{(y-y_1)(y-y_2)}{(y_0-y_1)(y_0-y_2)} + x_1\frac{(y-y_0)(y-y_2)}{(y_1-y_0)(y_1-y_2)} + x_2\frac{(y-y_0)(y-y_1)}{(y_2-y_0)(y_2-y_1)}$$
$$= \frac{1}{2}\frac{(y+1)(y-e+2)}{(\sqrt{e}-2+1)(\sqrt{e}-2-e+2)} + \frac{(y+1)(y-\sqrt{e}+2)}{(e-2+1)(e-2-\sqrt{e}+2)}.$$

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 connait pas de formule générale pour le calcul des zéros d'un polynôme de degré $n \geq 3$.

Interpolation trigonométrique

☆ Exercice 3.16
Considérons la fonction $f\colon [0;2\pi] \to \mathbb{R}$ 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 $x_j = j\pi/5$, $j = 0,\ldots,9$ à l'aide des instructions suivantes

```
% fonction a 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 calculee par Octave
z = interpft (Py ,S);
```

```
% tilde_f calculee 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');
```

CHAPITRE 4

De l'interpolation à l'approximation d'intégrales : formules de quadrature interpolatoires

4.1. Calcul analytique de primitives et intégrales

Une fonction $F\colon [a;b] \to \mathbb{R}$ est une PRIMITIVE (ou INTÉGRALE INDÉFINIE) d'une fonction $f\colon [a;b] \to \mathbb{R}$ si

$$F'(x) = f(x) \quad \text{pour tout } x \in [a;b].$$

- ⋆ Si F existe on dit que f est INTÉGRABLE.
- ⋆ F est une primitive de f ssi la fonction $F + c$ l'est pour tout réel c.
- ⋆ L'ensemble des primitives d'une fonction f est noté $\int f(x)\,\mathrm{d}x$:

$$\int f(x)\,\mathrm{d}x = F(x) + c.$$

EXEMPLE
- ⋆ $(x^2)' = 2x$ donc $\int 2x\,\mathrm{d}x = x^2 + c$
- ⋆ $(\ln(x))' = \frac{1}{x}$ donc $\int \frac{1}{x}\,\mathrm{d}x = \ln(x) + c$
- ⋆ $(\sin(x))' = \cos(x)$ donc $\int \cos(x)\,\mathrm{d}x = \sin(x) + c$

Calcul de primitives

1. $\displaystyle\int x^n\,\mathrm{d}x = \frac{x^{n+1}}{n+1} + c$ pour $n \neq -1$
2. $\displaystyle\int \frac{1}{x}\,\mathrm{d}x = \ln(x) + c$
3. $\displaystyle\int e^x\,\mathrm{d}x = e^x + c$
4. $\displaystyle\int \sin(x)\,\mathrm{d}x = -\cos(x) + c$
5. $\displaystyle\int \cos(x)\,\mathrm{d}x = \sin(x) + c$
6. $\displaystyle\int \frac{1}{\cos^2(x)}\,\mathrm{d}x = \tan(x) + c$
7. $\displaystyle\int \frac{1}{\sin^2(x)}\,\mathrm{d}x = -\frac{1}{\tan(x)} + c$
8. $\displaystyle\int \frac{1}{\sqrt{1-x^2}}\,\mathrm{d}x = \arcsin(x) + c = -\arccos(x) + c$
9. $\displaystyle\int \frac{1}{1+x^2}\,\mathrm{d}x = \arctan(x) + c$

Techniques d'intégration

Linéarité :

$$\boxed{\int (f(x) + g(x))\,\mathrm{d}x = \int f(x)\,\mathrm{d}x + \int g(x)\,\mathrm{d}x} \quad \text{et} \quad \boxed{\int kf(x)\,\mathrm{d}x = k\int f(x)\,\mathrm{d}x}, \ k \in \mathbb{R}$$

Produit (= intégration par parties) :

$$\boxed{\int f(x)g'(x)\,\mathrm{d}x = f(x)g(x) - \int f'(x)g(x)\,\mathrm{d}x}$$

Astuce mnémotechnique : on se rappelle de la formule de dérivation d'un produit

$$\big(f(x)\times g(x)\big)' = f'(x)\times g(x) + f(x)\times g'(x),$$

en intégrant cette expression

$$\int \big(f(x)\times g(x)\big)'\,\mathrm{d}x = \int f'(x)\times g(x)\,\mathrm{d}x + \int f(x)\times g'(x)\,\mathrm{d}x,$$

on obtient

$$f(x)\times g(x) = \int f'(x)\times g(x)\,\mathrm{d}x + \int f(x)\times g'(x)\,\mathrm{d}x,$$

Composition (=intégration par changement de variable) : en posant $u = f(x)$ on obtient $\frac{\mathrm{d}u}{\mathrm{d}x} = f'(x)$, soit encore $\mathrm{d}u = f'(x)\,\mathrm{d}x$ et donc

$$\int g(\underbrace{f(x)}_{=u})\,\underbrace{f'(x)\,\mathrm{d}x}_{=\mathrm{d}u} = \int g(u)\,\mathrm{d}u = G(u) + c = G(f(x)) + c$$

Astuce mnémotechnique : soit $G' = g$ alors

$$\Big(G\big(f(x)\big)\Big)' = G'\big(f(x)\big)\times f'(x),$$

en intégrant cette expression

$$\int \Big(G\big(f(x)\big)\Big)'\,\mathrm{d}x = \int G'\big(f(x)\big)\times f'(x)\,\mathrm{d}x,$$

on obtient

$$G\big(f(x)\big) = \int G'\big(f(x)\big)\times f'(x)\,\mathrm{d}x,$$

EXEMPLE

Voyons un exemple pour chaque méthode d'intégration :

1. $\displaystyle\int x(1+x)^3\,\mathrm{d}x = \int x(1+3x+3x^2+x^3)\,\mathrm{d}x = \int x+3x^2+3x^3+x^4\,\mathrm{d}x = \frac{x^2}{2}+3\frac{x^3}{3}+3\frac{x^4}{4}+\frac{x^5}{5}+c$
2. $\displaystyle\int \underbrace{(x^2+1)}_{u}{}^{50}\,\underbrace{2x\,\mathrm{d}x}_{\mathrm{d}u} = \int u^{50}\,\mathrm{d}u = \frac{u^{51}}{51}+c = \frac{(x^2+1)^{51}}{51}+c$
3. $\displaystyle\int \underbrace{x}_{f}\,\underbrace{e^{2x}}_{g'}\,\mathrm{d}x = \underbrace{x}_{f}\,\underbrace{\tfrac{1}{2}e^{2x}}_{g} - \int \underbrace{1}_{f'}\,\underbrace{\tfrac{1}{2}e^{2x}}_{g}\,\mathrm{d}x = \tfrac{1}{2}xe^{2x} - e^{2x} + c = \big(\tfrac{1}{2}x-1\big)e^{2x}+c$

EXEMPLE

L'intégration par changement de variables permet de calculer simplement les primitives suivantes :

1. $\displaystyle\int [u(x)]^n u'(x)\,\mathrm{d}x$ pour $n\neq -1$
2. $\displaystyle\int \frac{u'(x)}{u(x)}\,\mathrm{d}x$
3. $\displaystyle\int e^{u(x)}u'(x)\,\mathrm{d}x$
4. $\displaystyle\int \sin(u(x))u'(x)\,\mathrm{d}x$
5. $\displaystyle\int \cos(u(x))u'(x)\,\mathrm{d}x$
6. $\displaystyle\int \frac{u'(x)}{\cos^2(u(x))}\,\mathrm{d}x$
7. $\displaystyle\int \frac{u'(x)}{\sin^2(u(x))}\,\mathrm{d}x$
8. $\displaystyle\int \frac{u'(x)}{\sqrt{1-(u(x))^2}}\,\mathrm{d}x$
9. $\displaystyle\int \frac{u'(x)}{1+(u(x))^2}\,\mathrm{d}x$

En effet, il suffit de remplacer $u(x)$ par t (ainsi $u'(x)\,\mathrm{d}x = \mathrm{d}t$) :

1. $\displaystyle\int t^n\,\mathrm{d}t = \frac{[u(x)]^{n+1}}{n+1}+c$
2. $\displaystyle\int \frac{1}{t}\,\mathrm{d}t = \ln(|u(x)|)+c$
3. $\displaystyle\int e^t\,\mathrm{d}t = e^x + c$
4. $\displaystyle\int \sin(t)\,\mathrm{d}t = -\cos(u(x))+c$
5. $\displaystyle\int \cos(t)\,\mathrm{d}t = \sin(u(x))+c$
6. $\displaystyle\int \frac{1}{\cos^2(t)}\,\mathrm{d}t = \tan(u(x))+c$

7. $\int \frac{1}{\sin^2(t)}\,\mathrm{d}t = -\frac{1}{\tan(u(x))} + c$
8. $\int \frac{1}{1+t^2}\,\mathrm{d}t = \arctan(u(x)) + c$
9. $\int \frac{1}{\sqrt{1-t^2}}\,\mathrm{d}t = \arcsin(u(x)) + c = -\arccos(u(x) + c$

EXEMPLE
Calculer une primitive de $\frac{\ln(x)}{x}$ avec les trois méthodes.

- ⋆ INTÉGRATION DIRECTE :
$$\int \underbrace{\ln(x)}_{u(x)} \underbrace{\frac{1}{x}}_{u'(x)}\,\mathrm{d}x = \int u(x)u'(x)\,\mathrm{d}x = \frac{[u(x)]^2}{2} = \frac{\ln^2(x)}{2} + c$$
- ⋆ INTÉGRATION PAR CHANGEMENT DE VARIABLE :
$$\int \frac{\ln(x)}{x}\,\mathrm{d}x \underset{\substack{\uparrow \\ u=\ln(x) \\ e^u=x \\ e^u\,\mathrm{d}u=\mathrm{d}x}}{=} \int \frac{u}{e^u}e^u\,\mathrm{d}u = \int u\,\mathrm{d}u = \frac{u^2}{2} + c = \frac{\ln^2(x)}{2} + c$$
- ⋆ INTÉGRATION PAR PARTIES :
$$\int \underbrace{\ln(x)}_{f(x)} \underbrace{\frac{1}{x}}_{g'(x)}\,\mathrm{d}x \underset{\substack{\uparrow \\ f(x)=\ln(x)\Longrightarrow f'(x)=\frac{1}{x} \\ g'(x)=\frac{1}{x}\Longrightarrow g(x)=\ln(x)}}{=} \int f(x)g'(x)\,\mathrm{d}x = f(x)g(x) - \int f'(x)g(x)\,\mathrm{d}x = \ln^2(x) - \int \frac{\ln(x)}{x}\,\mathrm{d}x$$

i.e. $2\int \frac{\ln(x)}{x}\,\mathrm{d}x = \ln^2(x) + k$ et finalement $\int \frac{\ln(x)}{x}\,\mathrm{d}x = \frac{\ln^2(x)}{2} + c$.

4.1.1. Intégrale définie et interprétation géométrique

L'INTÉGRALE DÉFINIE sur $[a;b]$ d'une fonction $f\colon [a;b] \to \mathbb{R}$ est

$$\boxed{\int_a^b f(x)\,\mathrm{d}x = [F(x)]_a^b = F(b) - F(a)}$$

où $F'(x) = f(x)$ (*i.e.* F est une primitive de f) et l'on a les propriétés suivantes :

1. $\int_a^a f(x)\,\mathrm{d}x = 0$
2. $\int_a^b f(x)\,\mathrm{d}x = -\int_b^a f(x)\,\mathrm{d}x$
3. $\int_a^c f(x)\,\mathrm{d}x + \int_c^b f(x)\,\mathrm{d}x = \int_a^b f(x)\,\mathrm{d}x$ (Relation de Chasles)

EXEMPLE
- ⋆ $\int_1^2 x^2\,\mathrm{d}x = \left[\frac{x^3}{3}\right]_1^2 = \frac{2^3}{3} - \frac{1^3}{3} = \frac{7}{3}$
- ⋆ $\int_1^e \frac{1}{x}\,\mathrm{d}x = [\ln(x)]_1^e = \ln(e) - \ln(1) = 1$

EXEMPLE
Pour calculer l'intégrale d'une fonction définie par morceaux, on peut utiliser l'additivité de l'intégrale en sommant des intégrales où la fonction est définie simplement. Soit par exemple $f(x) = \begin{cases} 1, & \text{si } x < 1, \\ x, & \text{si } x \geq 1, \end{cases}$ alors

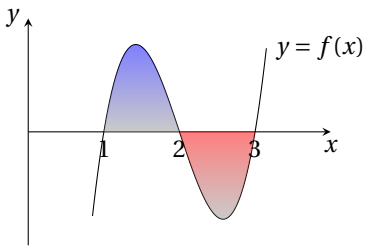
$$\int_0^3 f(x)\,\mathrm{d}x = \int_0^1 1\,\mathrm{d}x + \int_1^3 x\,\mathrm{d}x = [x]_0^1 + \left[\frac{x^2}{2}\right]_1^3 = 1 - 0 + \frac{3^2}{2} - \frac{1^2}{2} = 5$$

Un autre exemple : sachant que $|x| = \begin{cases} x & \text{si } x \geq 0 \\ -x & \text{si } x < 0 \end{cases}$ on a

$$\begin{aligned}\int_{-1}^{2} |x|\,\mathrm{d}x &= \int_{-1}^{0} |x|\,\mathrm{d}x + \int_{0}^{2} |x|\,\mathrm{d}x = \int_{-1}^{0} -x\,\mathrm{d}x + \int_{0}^{2} x\,\mathrm{d}x = -\int_{-1}^{0} x\,\mathrm{d}x + \int_{0}^{2} x\,\mathrm{d}x \\ &= -\left[\frac{x^2}{2}\right]_{-1}^{0} + \left[\frac{x^2}{2}\right]_{0}^{2} = -\left(\frac{0^2}{2} - \frac{(-1)^2}{2}\right) + \frac{2^2}{2} - \frac{0^2}{2} = -\left(0 - \frac{1}{2}\right) + \left(\frac{4}{2} - 0\right) = \frac{5}{2}\end{aligned}$$

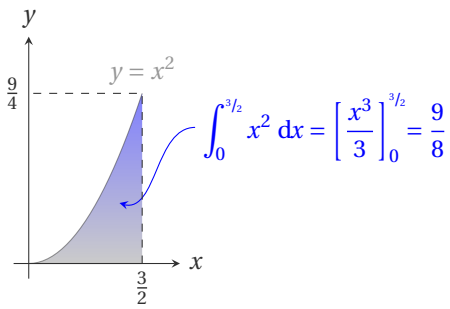
Interprétation géométrique : aire, déplacement, vitesse, accélération

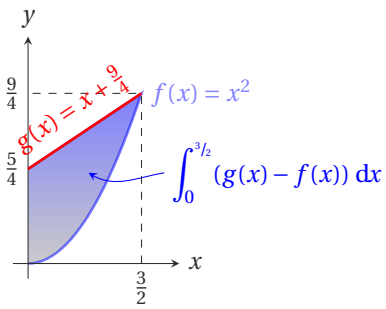
$$\int_a^b f(x)\,\mathrm{d}x = \text{(Aire au-dessus de l'axe des abscisses)} - \text{(Aire en dessous de l'axe des abscisses)}$$



- ⋆ Aire $= \int_1^2 f(x)\,\mathrm{d}x$
- ⋆ Aire $= -\int_2^3 f(x)\,\mathrm{d}x$

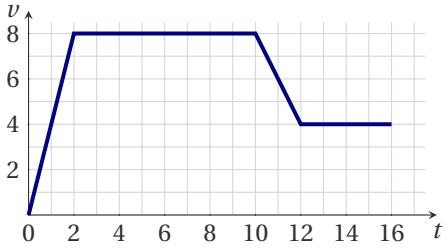
EXEMPLE





EXEMPLE

Celui ci-dessous est le graphe de la vitesse d'un coureur en fonction du temps. Quelle distance a-t-il parcouru en 16 s ?



La vitesse v en fonction du temps t est la dérivée du déplacement s.
Calculer un déplacement en connaissant la vitesse, signifie chercher une fonction $t \mapsto s(t)$ telle que $s'(t) = v(t)$, autrement dit, s est une primitive de v.

La distance parcourue sera donnée par $s(t_{\text{final}}) - s(t_{\text{initial}})$, autrement dit par $\int_{t_{\text{initial}}}^{t_{\text{final}}} v(t)\,\mathrm{d}t$:

$$\int_0^{16} v(t)\,\mathrm{d}t = \text{Aire sous la courbe} = 100\,\text{m}$$

EXEMPLE

1. Un train part de la station A et arrive à la station B en 6 min. Si la vitesse du train en mètres par minute est
$$v(t) = 24t^2(6-t),$$
quelle est la distance entre A et B ?
$$\int_0^6 v(t)\,\mathrm{d}t = \left[48t^3 - 6t^4\right]_0^6 = 2592\,\text{m}$$
2. Une particule a une accélération de $a(t) = 2t$. Si sa vitesse à l'instant $t = 1$ est $v(1) = 6$ et la distance parcourue à l'instant $t = 1$ depuis le point initial est $s(1) = 17$, quelle distance aura-t-elle parcourue à l'instant $t = 2$?
$a(t) = v'(t) = 2t$ donc $v(t) = t^2 + c_1$. Puisque $v(1) = 6$ alors $c_1 = 5$ et $v(t) = t^2 + 5$.
$v(t) = s'(t)$ donc $s(t) = \frac{t^3}{3} + 5t + c_2$. Puisque $s(1) = 17$ alors $c_2 = \frac{35}{3}$ et $s(t) = \frac{t^3}{3} + 5t + \frac{35}{3}$.
Par conséquent, $s(2) = \frac{73}{3}$.

4.2. Calcul approché d'intégrales

Pour une fonction arbitraire, il n'est pas toujours possible 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 ?

$$\begin{aligned}\text{erf}\colon \mathbb{R} &\to \mathbb{R}\\ x &\mapsto \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\,\mathrm{d}t\end{aligned}$$

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

$$\int_0^\pi f(x)\,\mathrm{d}x = \pi\frac{81}{16}\sum_{k=0}^{\infty}\frac{(-9/4)^k}{k!(k+4)!}.$$

On voit que le calcul de l'intégrale est transformé en un calcul, aussi difficile, de la somme d'une série. De plus, 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.

4.2.1. Principes généraux

Soit f une fonction réelle intégrable sur l'intervalle $[a; b]$. Le calcul explicite de l'intégrale définie $I_{[a;b]}(f) \equiv \int_a^b f(x)\,\mathrm{d}x$ peut être difficile, voire impossible.
On appelle *formule de quadrature* ou *formule d'intégration numérique* toute formule permettant de calculer une approximation de $I_{[a;b]}(f)$. Par exemple, on peut remplacer f par une approximation $\tilde{f}$ et calculer $I_{[a;b]}(\tilde{f})$ au lieu de $I_{[a;b]}(f)$:

$$\underbrace{\int_a^b f(x)\,\mathrm{d}x}_{I_{[a;b]}(f)} \approx \underbrace{\int_a^b \tilde{f}(x)\,\mathrm{d}x}_{I_{[a;b]}(\tilde{f})}$$

Pour améliorer la précision, on pourra utiliser l'additivité des intégrales et utiliser la formule de quadrature pour approcher chaque intégrale :

$$\int_a^b f(x)\,\mathrm{d}x = \int_a^c f(x)\,\mathrm{d}x + \int_c^b f(x)\,\mathrm{d}x \approx \int_a^c \tilde{f}_1(x)\,\mathrm{d}x + \int_c^b \tilde{f}_2(x)\,\mathrm{d}x$$

où $\tilde{f}_1$ est une approximation de f sur $[a;c]$ et $\tilde{f}_2$ est une approximation de f sur $[c;b]$. Cela amène aux formules de quadratures 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 de quadrature (généralement la même formule sur chaque sous-intervalle). Puisque

$$I_{[a;b]}(f) = \int_a^b f(x)\,\mathrm{d}x = \sum_{j=0}^{m-1}\int_{y_j}^{y_{j+1}} f(x)\,\mathrm{d}x = \sum_{j=0}^{m-1} I_{[y_j;y_{j+1}]}(f) \approx \sum_{j=0}^{m-1} I^{(j)}_{[y_j;y_{j+1}]}(\tilde{f}).$$

4.2.2. Exemples de formules de quadrature interpolatoires

L'approximation $\tilde{f}$ doit être facilement intégrable. Une approche naturelle consiste à prendre pour $\tilde{f}$ le polynôme p_n qui interpole f sur un ensemble de $n+1$ nœuds distincts $\{x_i\}_{i=0}^{i=n+1}$:

$$\underbrace{\int_a^b f(x)\,\mathrm{d}x}_{I_{[a;b]}(f)} \simeq \underbrace{\int_a^b p_n(x)\,\mathrm{d}x}_{I_{[a;b]}(p_n)}.$$

On sait qu'augmenter le nombre des points où on interpole f pour calculer $\tilde{f}$ ne garantit pas qu'on améliore l'approximation de f (se rappeler du phénomène de Runge). On utilisera donc des polynômes de degré petit (maximum 2) et pour améliorer la précision on utilisera des formules composites.

① La formule du *rectangle à gauche* est obtenue en remplaçant f par le polynôme qui interpole f en le point $(a, f(a))$ ce qui donne

$$p(x) = f(a)$$
$$I_{[a;b]}(f) \approx I_{[a;b]}(p) = (b-a)f(a).$$

② La formule du *rectangle à droite* est obtenue en remplaçant f par le polynôme qui interpole f en le point $(b, f(b))$ ce qui donne

$$p(x) = f(b)$$
$$I_{[a;b]}(f) \approx I_{[a;b]}(p) = (b-a)f(b).$$

③ La formule du *rectangle* ou du *point milieu* est obtenue en remplaçant f par le polynôme qui interpole f en le point $\left(\frac{a+b}{2}, f\left(\frac{a+b}{2}\right)\right)$ ce qui donne

$$p(x) = f\left(\tfrac{a+b}{2}\right)$$
$$I_{[a;b]}(f) \approx I_{[a;b]}(p) = (b-a)f\left(\tfrac{a+b}{2}\right).$$

④ La formule du *trapèze* est obtenue en remplaçant f par le polynôme qui interpole f en les points $(a, f(a))$ et $(b, f(b))$ ce qui donne

$$p(x) = \tfrac{f(b)-f(a)}{b-a}(x-a) + f(a)$$
$$I_{[a;b]}(f) \approx I_{[a;b]}(p) = \frac{b-a}{2}\big(f(a)+f(b)\big).$$

Nom	Points d'interpolation	Polynôme $p_n \in \mathbb{R}_n[t]$	$\int_a^b p_n(t)\,\mathrm{d}t$	
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	$p_1(t) = \frac{f(b)-f(a)}{b-a}(t-a) + f(a) \in \mathbb{R}_1[t]$	$\frac{(b-a)}{2}(f(a)+f(b))$	

Toutes ces formules peuvent bien sûr se généraliser pour obtenir des formules de quadrature composites. Pour chaque méthode, nous écrirons aussi une `function` qui prend en entrée `a`, `b` (avec $a < b$), `f` la fonction à intégrer et `m` (≥ 1) le nombre de sous-intervalles et donne en sortie `Int` la valeur de l'intégrale (et on ajoute un test).

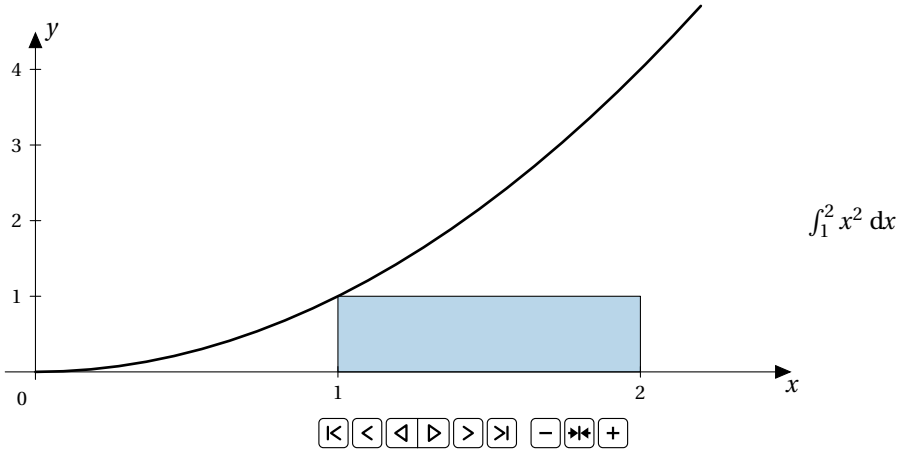
Formule du *rectangle à gauche*. La formule du *rectangle à gauche* est obtenue en remplaçant f par le polynôme qui interpole f en le point $(a, f(a))$, ce qui donne

$$p(x) = f(a) \quad \forall x \in [a;b]$$
$$\int_a^b f(x)\,\mathrm{d}x \approx \int_a^b p(x)\,\mathrm{d}x = (b-a)f(a).$$

Pour obtenir la formule 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 nœuds de quadrature $x_k = a + kH$ pour $k = 0, 1, \ldots, m-1$ on obtient la *formule composite du rectangle à gauche*

$$\int_a^b f(x)\,\mathrm{d}x = \sum_{k=0}^{m-1} \int_{x_k}^{x_{k+1}} f(x)\,\mathrm{d}x \simeq \sum_{k=0}^{m-1} \int_{x_k}^{x_{k+1}} p(x)\,\mathrm{d}x = H \sum_{k=0}^{m-1} f(x_k) = H \sum_{k=0}^{m-1} f(a+kH).$$

```
1;

function [Int]=gauche(a,b,f,m)
   h=(b-a)/m;
   x=[a:h:b-h];
   y=f(x);
   Int=h*sum( y );
   % en une ligne:
   % Int=(b-a)/m*sum(f([a:(b-a)/m:b](1:end-1)));
end

% TEST :
f=@(x)[3*x.^2];
gauche(0,1,f,10000)
% La valeur exacte est = 1
```



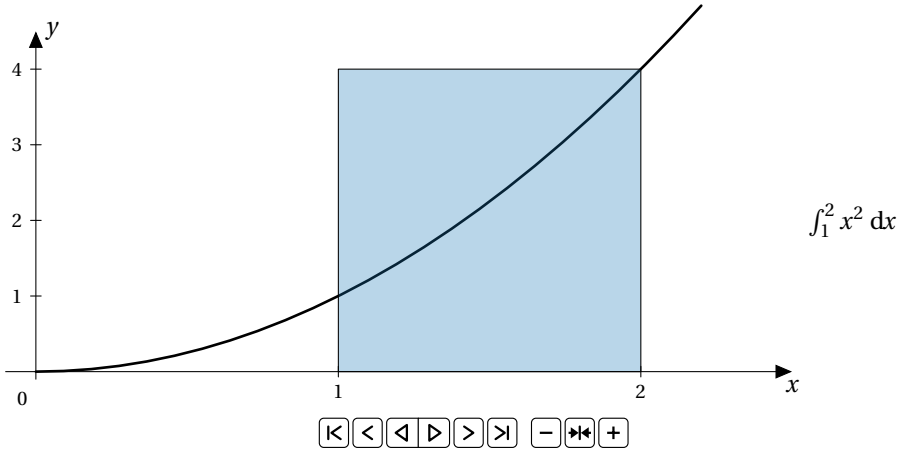
Formule du *rectangle à droite*. La formule du *rectangle à droite* est obtenue en remplaçant f par le polynôme qui interpole f en le point $(b, f(b))$, ce qui donne

$$p(x) = f(b) \quad \forall x \in [a;b]$$
$$\int_a^b f(x)\,\mathrm{d}x \approx \int_a^b p(x)\,\mathrm{d}x = (b-a)f(b).$$

Pour obtenir la formule 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 nœuds de quadrature $x_k = a + (k+1)H$ pour $k = 0, 1, \ldots, m-1$ on obtient la *formule composite du rectangle à droite*

$$\int_a^b f(x)\,\mathrm{d}x = \sum_{k=0}^{m-1} \int_{x_k}^{x_{k+1}} f(x)\,\mathrm{d}x \simeq \sum_{k=0}^{m-1} \int_{x_k}^{x_{k+1}} p(x)\,\mathrm{d}x = H \sum_{k=0}^{m-1} f(x_{k+1}) = H \sum_{k=0}^{m-1} f(a+(k+1)H).$$


```
1;

function [Int]=droite(a,b,f,m)
   h=(b-a)/m;
   x=[a+h:h:b];
   y=f(x);
   Int=h*sum( y );
end

% TEST :
f=@(x)[3*x.^2];
droite(0,1,f,10000)
% La valeur exacte est = 1
```



$\int_1^2 x^2 \, \mathrm{d}x$

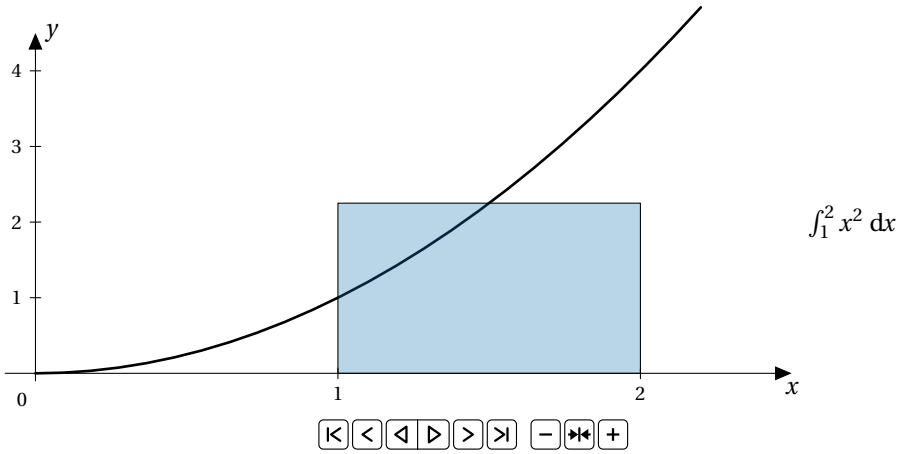
Formule du *rectangle* ou du *point milieu*. La formule du *rectangle* ou du *point milieu* est obtenue en remplaçant f par une le polynôme qui interpole f en le point $\left(\frac{a+b}{2}, f\left(\frac{a+b}{2}\right)\right)$, ce qui donne

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

Pour obtenir la formule composite 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 $x_k = a + \frac{H}{2} + kH$ pour $k = 0, 1, \ldots, m-1$ on obtient la *formule composite du point milieu*

$$\int_a^b f(x) = \sum_{k=0}^{m-1} \int_{x_k}^{x_{k+1}} f(x)\,\mathrm{d}x \simeq \sum_{k=0}^{m-1} \int_{x_k}^{x_{k+1}} p(x)\,\mathrm{d}x = H \sum_{k=0}^{m-1} f(x_k) = H \sum_{k=0}^{m-1} f\left(a + \frac{H}{2} + kH\right).$$

```
1;

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

% TEST :
f=@(x)[3*x.^2];
milieu(0,1,f,10000)
% La valeur exacte est = 1
```



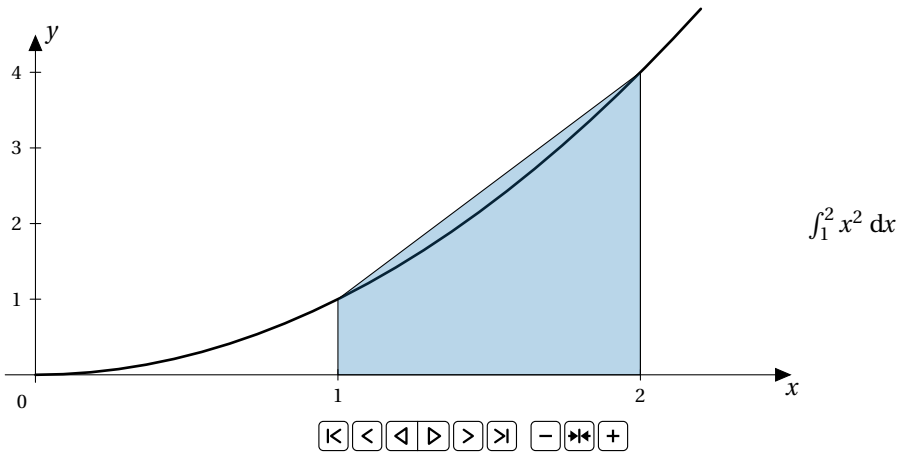
Formule du *trapèze*. La formule du *trapèze* est obtenue en remplaçant f par le segment qui relie $(a, f(a))$ à $(b, f(b))$, *i.e.* le polynôme qui interpole f en les points $(a, f(a))$ et $(b, f(b))$, ce qui donne

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

Pour obtenir la formule composite 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 $x_k = a + kH$ pour $k = 0, 1, \ldots, m-1$ on obtient la *formule composite des trapèzes*

$$\begin{aligned}\int_a^b f(x) &= \sum_{k=0}^{m-1} \int_{x_k}^{x_{k+1}} f(x)\,\mathrm{d}x \simeq \sum_{k=0}^{m-1} \int_{x_k}^{x_{k+1}} p(x)\,\mathrm{d}x \\ &= \frac{H}{2}\sum_{k=0}^{m-1}\big(f(x_k)+f(x_{k+1})\big) = H\left(\frac{1}{2}f(a) + \sum_{k=1}^{m-1} f(a+kH) + \frac{1}{2}f(b)\right).\end{aligned}$$

```
1;

function [Int]=trapeze(a,b,f,m)
   h=(b-a)/m;
   x=[a:h:b];
   y=f(x);
   Int=h*( 0.5*y(1)+sum(y(2:end-1))+0.5*y(end) );
end

% TEST :
f=@(x)[3*x.^2];
trapeze(0,1,f,10000)
% La valeur exacte est = 1
```



4.3. Exercices

4.3.1. Calcul analytique de primitives et intégrales

Primitives : techniques d'intégration

Exercice 4.1 (Par intégration directe)
Calculer les primitives suivantes :

1. $\int 2x^3 - 3x + 1 \, dx$
2. $\int \sqrt{x} + \sqrt[3]{x} \, dx$
3. $\int \frac{1}{\sqrt{x+1}} \, dx$
4. $\int \sqrt[4]{(x-1)^3} \, dx$
5. $\int \frac{1}{x\sqrt[3]{x}} \, dx$
6. $\int (1+2x^3)^2 \, dx$
7. $\int \frac{x}{x+1} \, dx$
8. $\int \frac{x^3+x+1}{x^2+1} \, dx$

Correction

1. $2\frac{x^4}{4} - 3\frac{x^2}{2} + x + c = \frac{1}{2}x(x^3 - 3x + 2) + c$
2. $\frac{x^{3/2}}{3/2} + \frac{x^{4/3}}{4/3} + c$
3. $\frac{(x+1)^{1/2}}{1/2} + c = 2\sqrt{x+1} + c$
4. $\frac{4}{7}(x-1)^{7/4} + c$
5. $-\frac{3}{\sqrt[3]{x}} + c$
6. $\frac{4}{7}x^7 + x^4 + x + c$
7. $\int \frac{x+1-1}{x+1} \, dx = \int 1 - \frac{1}{x+1} \, dx = x - \ln(x+1) + c$
8. $\int \frac{x(x^2+1)+1}{x^2+1} \, dx = \int x + \frac{1}{x^2+1} \, dx = \frac{x^2}{2} + \arctan(x) + c$

Exercice 4.2
Calculer les primitives suivantes :

1. $\int \frac{1}{1+x} \, dx$
2. $\int \frac{1}{1-x} \, dx$
3. $\int \frac{1}{1+x^2} \, dx$
4. $\int \frac{1}{1-x^2} \, dx$

Correction

1. $u(x) = 1 + x$ et $u'(x) = 1$ donc $\int \frac{u'(x)}{u(x)} \, dx = \ln|1+x| + c$
2. $u(x) = 1 - x$ et $u'(x) = -1$ donc $-\int \frac{u'(x)}{u(x)} \, dx = -\ln|1-x| + c$
3. $\arctan(x) + c$
4. $\frac{1}{1-x^2} = \frac{1/2}{1-x} + \frac{1/2}{1+x}$ donc $\frac{1}{2}\left(\int \frac{1}{1-x} \, dx + \int \frac{1}{1+x} \, dx\right) = \frac{1}{2}(\ln|1+x| - \ln|1-x|) + c = \frac{1}{2}\ln\left(\frac{1+x}{1-x}\right) + c$

Exercice 4.3 (Par transformations élémentaires)
Calculer les primitives suivantes en utilisant $\int f(u(x))u'(x) \, dx = \int f(t) \, dt$:

1. $\int \frac{e^x}{1+e^x} \, dx$
2. $\int \sqrt{2x+1} \, dx$
3. $\int \frac{1}{1+e^x} \, dx$
4. $\int \frac{\ln^3(x)}{x} \, dx$
5. $\int \frac{e^{-1/x}}{x^2} \, dx$
6. $\int \frac{1}{x\ln^3(x)} \, dx$
7. $\int \frac{1+\cos(x)}{x+\sin(x)} \, dx$
8. $\int \frac{2x}{1+x^4} \, dx$
9. $\int \frac{\sin(2x)}{1+\sin^2(x)} \, dx$
10. $\int \sin^3(x)\cos(x) \, dx$
11. $\int \frac{1}{\sin(x)\cos(x)} \, dx$
12. $\int \frac{e^{\tan(x)}}{\cos^2(x)} \, dx$
13. $\int \frac{1-2x}{\sqrt{1-x^2}} \, dx$
14. $\int \frac{x+1}{x^2+2x+2} \, dx$
15. $\int x(x^2+1)^2 \, dx$
16. $\int e^{2x+1} \, dx$
17. $\int x\sqrt{5+x^2} \, dx$
18. $\int \frac{e^{\sqrt{x}}}{\sqrt{x}} \, dx$
19. $\int xe^{x^2} \, dx$
20. $\int x^2 e^{x^3} \, dx$
21. $\int \frac{\sin(\sqrt{x})}{\sqrt{x}} \, dx$
22. $\int \frac{x}{\sqrt[3]{x^2+3}} \, dx$
23. $\int \frac{x^3}{1+x^4} \, dx$
24. $\int \sin(3x) \, dx$

Correction

1. $u(x)=1+e^x$, $u'(x)=e^x$, $\int \frac{u'(x)}{u(x)}\,\mathrm{d}x=\ln(1+e^x)+c$
2. $u(x)=2x+1$, $u'(x)=2$, $\frac{1}{2}\int\sqrt{u(x)}u'(x)\,\mathrm{d}x=\frac{1}{2}\int[u(x)]^{1/2}u'(x)\,\mathrm{d}x=\frac{[u(x)]^{3/2}}{3}=\frac{\sqrt{(2x+1)^3}}{3}+c$
3. $u(x)=1+e^x$, $u'(x)=e^x$, $\int\frac{1+e^x-e^x}{1+e^x}\,\mathrm{d}x=\int 1\,\mathrm{d}x-\int\frac{u'(x)}{u(x)}\,\mathrm{d}x=x-\ln(1+e^x)+c$
4. $u(x)=\ln(x)$, $u'(x)=1/x$, $\int(u(x))^3u'(x)\,\mathrm{d}x=\dfrac{\ln^4(x)}{4}+c$
5. $u(x)=-1/x$, $u'(x)=1/x^2$, $\int e^{u(x)}u'(x)\,\mathrm{d}x=e^{-1/x}+c$
6. $u(x)=\ln(x)$, $u'(x)=1/x$, $\int\frac{u'(x)}{(u(x))^3}\,\mathrm{d}x=-\frac{1}{2\ln^2(x)}+c$
7. $u(x)=x+\sin(x)$, $u'(x)=1+\cos(x)$, $\int\frac{u'(x)}{u(x)}\,\mathrm{d}x=\ln|x+\sin(x)|+c$
8. $u(x)=x^2$, $u'(x)=2x$, $\int\frac{u'(x)}{1+(u(x))^2}\,\mathrm{d}x=\arctan(x^2)+c$
9. $u(x)=1+\sin^2(x)$, $u'(x)=2\sin(x)\cos(x)=\sin(2x)$, $\int\frac{u'(x)}{u(x)}\,\mathrm{d}x=\ln(1+\sin^2(x))+c$
10. $u(x)=\sin(x)$, $u'(x)=\cos(x)$, $\int(u(x))^3u'(x)\,\mathrm{d}x=\frac{\sin^4(x)}{4}+c$
11. $u(x)=\tan(x)$, $u'(x)=1/\cos^2(x)$, $\int\frac{\cos(x)}{\sin(x)\cos^2(x)}\,\mathrm{d}x=\int\frac{u'(x)}{u(x)}\,\mathrm{d}x=\ln|\tan(x)|+c$
12. $u(x)=\tan(x)$, $u'(x)=1/\cos^2(x)$, $\int e^{u(x)}u'(x)\,\mathrm{d}x=e^{\tan(x)}+c$
13. $u(x)=1-x^2$, $u'(x)=-2x$, $\int\frac{1}{\sqrt{1-x^2}}\,\mathrm{d}x+\int(u(x))^{-1/2}u'(x)\,\mathrm{d}x=\arcsin(x)+2\sqrt{1-x^2}+c$
14. $u(x)=x^2+2x+1$, $u'(x)=2x+2$, $\frac{1}{2}\int\frac{u'(x)}{u(x)}\,\mathrm{d}x=\frac{1}{2}\ln(x^2+2x+1)+c$
15. $u(x)=x^2+1$, $u'(x)=2x$, $\frac{1}{2}\int(u(x))^2u'(x)\,\mathrm{d}x=\frac{(x^2+1)^3}{6}+c$
16. $u(x)=2x+1$, $u'(x)=2$, $\frac{1}{2}\int e^{u(x)}u'(x)\,\mathrm{d}x=\frac{1}{2}e^{u(x)}+c=\frac{e^{2x+1}}{2}+c$
17. $u(x)=5+x^2$, $u'(x)=2x$, $\frac{1}{2}\int(u(x))^{1/2}u'(x)\,\mathrm{d}x=\frac{1}{3}(5+x^2)^{3/2}+c$
18. $u(x)=\sqrt{x}$, $u'(x)=\frac{1}{2\sqrt{x}}$, $2\int e^{u(x)}u'(x)\,\mathrm{d}x=2e^{\sqrt{x}}+c$
19. $u(x)=x^2$, $u'(x)=2x$, $\frac{1}{2}\int e^{u(x)}u'(x)\,\mathrm{d}x=\frac{e^{x^2}}{2}+c$
20. $u(x)=x^3$, $u'(x)=3x^2$, $\frac{1}{3}\int e^{u(x)}u'(x)\,\mathrm{d}x=\frac{e^{x^3}}{3}+c$
21. $u(x)=\sqrt{x}$, $u'(x)=\frac{1}{2\sqrt{x}}$, $\frac{1}{2}\int\sin(u(x))\,u'(x)\,\mathrm{d}x=-2\cos\sqrt{x}+c$
22. $u(x)=x^2+3$, $u'(x)=2x$, $\frac{1}{2}\int(u(x))^{-1/3}u'(x)\,\mathrm{d}x=\dfrac{3}{4}\sqrt[3]{(x^2+3)^2}+c$
23. $u(x)=1+x^4$, $u'(x)=4x^3$, $\frac{1}{4}\int\dfrac{u'(x)}{u(x)}\,\mathrm{d}x=\frac{\ln(1+x^4)}{4}+c$
24. $u(x)=3x$, $u'(x)=3$, $\frac{1}{3}\int\sin(u(x))u'(x)\,\mathrm{d}x=-\frac{1}{3}\cos(3x)+c$

Exercice 4.4

Calculer les primitives suivantes :

1. $\displaystyle\int\frac{\cos^2(x)}{1-\sin(x)}\,\mathrm{d}x$
2. $\displaystyle\int\frac{\sin(x)+\cos(x)}{\sin(x)-\cos(x)}\,\mathrm{d}x$
3. $\displaystyle\int\frac{1}{\sin^2(x)\cos^2(x)}\,\mathrm{d}x$
4. $\displaystyle\int\frac{1}{\cos(x)\sin(x)}\,\mathrm{d}x$
5. $\displaystyle\int\frac{1}{\sin(x)}\,\mathrm{d}x$
6. $\displaystyle\int\frac{1}{\cos(x)}\,\mathrm{d}x$

Correction

1. $\cos^2(x)=1-\sin^2(x)=(1-\sin(x))(1+\sin(x))$, $\int\frac{(1-\sin(x))(1+\sin(x))}{1-\sin(x)}\,\mathrm{d}x=x-\cos(x)+c$
2. $1=\cos^2(x)+\sin^2(x)$, $u(x)=\sin(x)-\cos(x)$, $u'(x)=\sin(x)+\cos(x)$, $\int\frac{\sin(x)+\cos(x)}{\sin(x)-\cos(x)}\,\mathrm{d}x=\int\frac{u'(x)}{u(x)}\,\mathrm{d}x=\ln(\sin(x)-\cos(x))+c$
3. $1=\cos^2(x)+\sin^2(x)$, $\int\frac{\sin^2(x)+\cos^2(x)}{\sin^2(x)\cos^2(x)}\,\mathrm{d}x=\int\frac{1}{\cos^2(x)}\,\mathrm{d}x+\int\frac{1}{\sin^2(x)}\,\mathrm{d}x=\tan(x)-\dfrac{1}{\tan(x)}+c$
4. $1=\cos^2(x)+\sin^2(x)$, $\int\frac{1}{\cos(x)\sin(x)}\,\mathrm{d}x=\int\frac{\cos^2(x)+\sin^2(x)}{\cos(x)\sin(x)}\,\mathrm{d}x=\int\frac{\sin(x)}{\cos(x)}+\frac{\cos(x)}{\sin(x)}\,\mathrm{d}x=-\ln|\cos(x)|+\ln|\sin(x)|+c$
5. $\sin(x)=2\sin(x/2)\cos(x/2)$, $u(x)=x/2$, $u'(x)=1/2$, $\int\frac{1}{\sin(x)}\,\mathrm{d}x=\int\frac{1}{\cos(u(x))\sin(u(x))}u'(x)\,\mathrm{d}x=-\ln|\cos(x/2)|+\ln|\sin(x/2)|+c=\ln|\tan(x/2)|+c$.
6. $\cos(x)=\sin\left(\frac{\pi}{2}-x\right)$, $u(x)=\frac{\pi}{2}-x$, $u'(x)=-1$, $\int\frac{1}{\cos(x)}\,\mathrm{d}x=-\int\frac{1}{\sin\left(u(x)\right)}u'(x)\,\mathrm{d}x=\ln|\cos(\frac{\pi}{4}-\frac{x}{2})|-\ln|\sin(\frac{\pi}{4}-\frac{x}{2})|+c=\ln(\sin(x)-\cos(x))+c$

Exercice 4.5 (Intégration par changement de variable)
Calculer les primitives suivantes :

1. $\int \frac{\sin(\ln(x))}{x}\,\mathrm{d}x$
2. $\int \frac{1+e^{\sqrt{x}}}{\sqrt{x}}\,\mathrm{d}x$
3. $\int \frac{1}{x-\sqrt{x}}\,\mathrm{d}x$
4. $\int \frac{1}{\sqrt{x(1+\sqrt{x})}}\,\mathrm{d}x$
5. $\int \frac{e^x}{1+e^x}\,\mathrm{d}x$
6. $\int \frac{1}{x\ln(x)}\,\mathrm{d}x$
7. $\int \frac{1}{x\sqrt{\ln\left(\frac{1}{x}\right)}}\,\mathrm{d}x$
8. $\int e^x \ln(1+e^x)\,\mathrm{d}x$
9. $\int \frac{1}{x(2+\ln^2(x))}\,\mathrm{d}x$
10. $\int \frac{x^3}{\sqrt{1-x^2}}\,\mathrm{d}x$
11. $\int \frac{x^5}{\sqrt{x^3-1}}\,\mathrm{d}x$
12. $\int \sqrt{e^x-1}\,\mathrm{d}x$
13. $\int \frac{\ln(x)}{x}\,\mathrm{d}x$
14. $\int \frac{1}{e^x+e^{-x}}\,\mathrm{d}x$
15. $\int \frac{e^{\tan(x)}}{\cos^2(x)}\,\mathrm{d}x$
16. $\int \frac{x}{\sqrt{1+x^2}}\,\mathrm{d}x$
17. $\int x\sqrt{a+x^2}\,\mathrm{d}x$
18. $\int \frac{1}{x\sqrt{1-\ln^2(x)}}\,\mathrm{d}x$
19. $\int \frac{e^{1/x}}{x^2}\,\mathrm{d}x$
20. $\int \frac{\cos(x)}{1+\sin(x)}\,\mathrm{d}x$
21. $\int \frac{1}{x^2}\cos\left(\frac{1}{x}\right)\mathrm{d}x$
22. $\int \frac{x^3}{\sqrt{1+x^2}}\,\mathrm{d}x$
23. $\int \frac{1}{3+x^2}\,\mathrm{d}x$
24. $\int \frac{x}{1+x^4}\,\mathrm{d}x$

Correction

1. Pour $x>0$, si on pose $t=\ln(x)$ alors $\frac{1}{x}\,\mathrm{d}x=\,\mathrm{d}t$ et on obtient $-\cos(\ln(x))+c$
2. Pour $x>0$, si on pose $t=\sqrt{x}$ alors $\mathrm{d}x=2t\,\mathrm{d}t$ et on obtient $2(\sqrt{x}+e^{\sqrt{x}})+c$
3. Pour $x>0$, si on pose $t=\sqrt{x}$ alors $\mathrm{d}x=2t\,\mathrm{d}t$ et on obtient $2\ln(\sqrt{x}-1)+c$
4. Pour $x>0$, si $x>0$. Si on pose $t=\sqrt{x}$ alors $\mathrm{d}x=2t\,\mathrm{d}t$ et on obtient $4\sqrt{1+\sqrt{x}}+c$
5. Si on pose $t=e^x$ alors $e^x\,\mathrm{d}x=\,\mathrm{d}t$ et on obtient $\ln(1+e^x)+c$
6. Pour $x>0$, si on pose $t=\ln(x)$ alors $\frac{1}{x}\,\mathrm{d}x=\,\mathrm{d}t$ et on obtient $\ln|\ln(x)|+c$
7. Pour $x>0$, si on pose $t=\ln\left(\frac{1}{x}\right)=-\ln(x)$ alors $-\frac{1}{x}\,\mathrm{d}x=\,\mathrm{d}t$ et on obtient $-2\sqrt{\ln\left(\frac{1}{x}\right)}+c$
8. Si on pose $t=1+e^x$ alors $e^x\,\mathrm{d}x=\,\mathrm{d}t$ ainsi $\int e^x\ln(1+e^x)\,\mathrm{d}x=\int\ln(t)\,\mathrm{d}t$. Sans utiliser l'intégration par partie, si on pose $t=e^w$ alors $\mathrm{d}t=e^w\,\mathrm{d}w$ ainsi $\int\ln(t)\,\mathrm{d}t=\int we^w\,\mathrm{d}w=(w-1)e^w$ et on obtient $(1+e^x)\ln(1+e^x)-e^x+c$.
9. Pour $x>0$, si on pose $t=\ln(x)$ alors $\frac{1}{x}\,\mathrm{d}x=\,\mathrm{d}t$ et on obtient $\frac{1}{\sqrt{2}}\arctan\left(\frac{\ln(x)}{\sqrt{2}}\right)+c$
10. Si on pose $t^2=1-x^2$ alors $-x\,\mathrm{d}x=t\,\mathrm{d}t$ et on obtient $-\frac{1}{3}(x^2+2)\sqrt{1-x^2}+c$
11. Si on pose $t^2=x^3-1$ alors $3x^2\,\mathrm{d}x=2t\,\mathrm{d}t$ et on obtient $\frac{2}{9}(x^3+2)\sqrt{x^3-1}+c$
12. Si on pose $t^2=e^x-1$ alors $\mathrm{d}x=\frac{2t}{t^2+1}\,\mathrm{d}t$ et on obtient $2\left(\sqrt{e^x-1}-\arctan(\sqrt{e^x-1})\right)+c$
13. Pour $x>0$, si on pose $t=\ln(x)$ alors $\mathrm{d}x=e^t\,\mathrm{d}t$ et on obtient $\frac{1}{2}\ln^2(x)+c$
14. Si on pose $t=e^x$ alors $\mathrm{d}x=\frac{1}{t}\,\mathrm{d}t$ et on obtient $\arctan e^x+c$
15. Si on pose $t=\tan(x)$ alors $\mathrm{d}x=\frac{1}{1+t^2}\,\mathrm{d}t$ et on obtient $e^{\tan(x)}+c$
16. Si on pose $t=\sqrt{1+x^2}$ alors $2x\,\mathrm{d}x=2t\,\mathrm{d}t$ et on obtient $\sqrt{1+x^2}+c$
17. Pour $a+x^2\geq 0$, si on pose $t=\sqrt{a+x^2}$ alors $2x\,\mathrm{d}x=2t\,\mathrm{d}t$ et on obtient $\frac{1}{3}\sqrt{(a+x^2)^3}+c$
18. Pour $x>0$, si on pose $t=\ln(x)$ alors $\frac{1}{x}\,\mathrm{d}x=\,\mathrm{d}t$ et on obtient $\arcsin(\ln(x))+c$
19. Si on pose $t=\frac{1}{x}$ alors $-\frac{1}{x^2}\,\mathrm{d}x=\,\mathrm{d}t$ et on obtient $-e^{1/x}+c$
20. Si on pose $t=1+\sin(x)$ alors $\cos(x)\,\mathrm{d}x=\,\mathrm{d}t$ et on obtient $\ln|1+\sin(x)|+c$
21. Si on pose $t=\frac{1}{x}$ alors $-\frac{1}{x^2}\,\mathrm{d}x=\,\mathrm{d}t$ et on obtient $-\sin\left(\frac{1}{x}\right)+c$
22. Si on pose $t^2=1+x^2$ alors $x\,\mathrm{d}x=t\,\mathrm{d}t$ et on obtient $\frac{1}{3}(x^2-2)\sqrt{x^2+1}+c$
23. $\int\frac{1}{3+x^2}\,\mathrm{d}x=\frac{1}{3}\int\frac{1}{1+\left(\frac{x}{\sqrt{3}}\right)^2}\,\mathrm{d}x$. Si on pose $t=x/\sqrt{3}$ alors $\mathrm{d}x=\sqrt{3}t\,\mathrm{d}t$ et on obtient $\frac{\sqrt{3}}{3}\arctan(x/\sqrt{3})+c$
24. Si on pose $t=x^2$ alors $2\,\mathrm{d}x=\,\mathrm{d}t$ et on obtient $\frac{1}{2}\arctan(x^2)+c$

Exercice 4.6 (Intégration par parties)
Calculer les primitives suivantes en utilisant $\int f(x)g'(x)\,dx = f(x)g(x) - \int f'(x)g(x)\,dx$:

1. $\int \frac{\ln(x)}{x^2}\,dx$
2. $\int \ln(1+x)\,dx$
3. $\int x^2 e^x\,dx$
4. $\int \frac{\ln(x)}{\sqrt{x}}\,dx$
5. $\int x\sin(x)\,dx$
6. $\int x\ln(x)\,dx$
7. $\int x^2\cos(x)\,dx$
8. $\int \frac{\sin(x)}{\cos^3(x)} e^{\tan(x)}\,dx$
9. $\int x^3\ln(x)\,dx$
10. $\int \frac{\ln(x)}{\sqrt[4]{x}}\,dx$
11. $\int \ln^2(x)\,dx$
12. $\int x^3\sin(x^2)\,dx$

Correction

1. On pose $f(x) = \ln(x)$ et $g'(x) = \frac{1}{x^2}$. Alors $f'(x) = \frac{1}{x}$ et $g(x) = -\frac{1}{x}$. On obtient $f(x)g(x) - \int f'(x)g(x)\,dx = -\frac{1+\ln(x)}{x} + c$
2. On pose $f(x) = \ln(1+x)$ et $g'(x) = 1$. Alors $f'(x) = \frac{1}{1+x}$ et $g(x) = x$. On obtient $f(x)g(x) - \int f'(x)g(x)\,dx = (1+x)\ln(1+x) - x + c$
3. On pose $f(x) = x^2$ et $g'(x) = e^x$. Alors $f'(x) = 2x$ et $g(x) = e^x$. On obtient $f(x)g(x) - \int f'(x)g(x)\,dx = e^x((x-2)x+2) + c$
4. On pose $f(x) = \ln(x)$ et $g'(x) = \frac{1}{\sqrt{x}}$. Alors $f'(x) = \frac{1}{x}$ et $g(x) = 2\sqrt{x}$. On obtient $f(x)g(x) - \int f'(x)g(x)\,dx = 2\sqrt{x}(\ln(x) - 2) + c$
5. On pose $f(x) = x$ et $g'(x) = \sin(x)$. Alors $f'(x) = 1$ et $g(x) = -\cos(x)$. On obtient $f(x)g(x) - \int f'(x)g(x)\,dx = -x\cos(x) + \sin(x) + c$
6. On pose $f(x) = \ln(x)$ et $g'(x) = x$. Alors $f'(x) = 1/x$ et $g(x) = x^2/2$. On obtient $f(x)g(x) - \int f'(x)g(x)\,dx = \frac{1}{2}x^2\ln(x) - \frac{1}{4}x^2 + c$
7. On pose $f(x) = x^2$ et $g'(x) = \cos(x)$. Alors $f'(x) = 2x$ et $g(x) = \sin(x)$. On obtient $f(x)g(x) - \int f'(x)g(x)\,dx = x^2\sin(x) - 2[-x\cos(x) + \sin(x)] + c$
8. On pose $f(x) = {}^{\sin(x)}/_{\cos(x)}$ et $g'(x) = {}^{e^{\tan(x)}}/_{x^2}$. Alors $f'(x) = {}^{1}/_{cos^2 x}$ et $g(x) = e^{\tan(x)}$. On obtient $f(x)g(x) - \int f'(x)g(x)\,dx = e^{\tan(x)}(\tan(x) - 1) + c$
9. On pose $f(x) = \ln(x)$ et $g'(x) = x^3$. Alors $f'(x) = \frac{1}{x}$ et $g(x) = \frac{x^4}{4}$. On obtient $f(x)g(x) - \int f'(x)g(x)\,dx = \frac{1}{16}x^4(4\ln(x) - 1) + c$
10. On pose $f(x) = \ln(x)$ et $g'(x) = \frac{1}{\sqrt[4]{x}}$. Alors $f'(x) = \frac{1}{x}$ et $g(x) = \frac{4x^{3/4}}{3}$. On obtient $f(x)g(x) - \int f'(x)g(x)\,dx = \frac{4}{3}x^{3/4}\left(\ln(x) - \frac{4}{3}\right) + c$
11. On pose $f(x) = \ln(x)$ et $g'(x) = \ln(x)$. Alors $f'(x) = \frac{1}{x}$ et $g(x) = x\ln(x) - x$. On obtient $f(x)g(x) - \int f'(x)g(x)\,dx = x(\ln^2(x) - 2\ln(x) + 2) + c$
12. On pose d'abord $t = x^2$ ainsi $dt = 2x\,dx$. Alors $\int x^3\sin(x^2)\,dx = \frac{1}{2}\int x^2\sin(x^2)2x\,dx = \frac{1}{2}\int t\sin(t)\,dt$. On pose $f(t) = t$ et $g'(t) = \sin(t)$. Alors $f'(t) = 1$ et $g(t) = -\cos(t)$. On obtient $f(t)g(t) - \int f'(t)g(t)\,dt = \frac{-t\cos(t) + \sin(t)}{2} + c = \frac{-x^2\cos(x^2) + \sin(x^2)}{2} + c$

Exercice 4.7
Calculer la primitive suivante en utilisant un changement de variable. Comparer ensuite au résultat obtenu en utilisant l'intégration par parties :

$$\int \frac{x\arcsin(x)}{\sqrt{1-x^2}}\,dx$$

Correction
CV Si on pose $t = \arcsin(x)$ alors $\frac{1}{\sqrt{1-x^2}}\,dx = dt$ et $x = \sin(t)$ et on obtient

$$\int \frac{x\arcsin(x)}{\sqrt{1-x^2}}\,dx = \int t\sin(t)\,dt$$

On a calculé cette intégrale à l'exercice 4.6(5.) :

$$\int t\sin(t)\,dt = -t\cos(t) + \sin(t) + c = -\arcsin(x)\cos(\arcsin(x)) + x + c$$

IPP

$$\int \frac{x\arcsin(x)}{\sqrt{1-x^2}}\,\mathrm{d}x \quad \begin{array}{ccc} & = & \\ f(x)=\arcsin(x) & \Rightarrow & f'(x)=\frac{1}{\sqrt{1-x^2}} \\ g(x)=-\sqrt{1-x^2} & \Leftarrow & g'(x)=\frac{x}{\sqrt{1-x^2}} \end{array} \quad -\sqrt{1-x^2}\cdot\arcsin(x)+x+c$$

Les deux calculs donnent le même résultat car $\cos(\arcsin(x))=\pm\sqrt{1-\sin^2(\arcsin(x))}=\pm\sqrt{1-x^2}$

Exercice 4.8 (cf. P. HALMOS)
Si $f,g\colon\mathbb{R}\to\mathbb{R}$ sont deux fonctions dérivables quelconques, on sait que **la dérivée du produit n'est pas le produit des dérivées**, autrement dit $(fg)'\neq f'g'$. Cependant, il existe des fonctions f et g pour lesquelles on a bien $(fg)'=f'g'$, par exemple si f et g sont toutes deux égales à une constante (pas nécessairement la même). Pouvez-vous en trouver d'autres ?

Correction

- ⋆ Si $f(x)=k_1$ pour tout $x\in\mathbb{R}$ et $g(x)=k_2$ pour tout $x\in\mathbb{R}$, alors $(fg)'(x)=(k_1k_2)'=0$ pour tout $x\in\mathbb{R}$ et $f'(x)g'(x)=0\times 0=0$ pour tout $x\in\mathbb{R}$.
- ⋆ Si $g=f$, on cherche f telle que $(f^2)'=(f')^2$, c'est-à-dire $2f(x)f'(x)=(f'(x))^2$ pour tout $x\in\mathbb{R}$. Donc, soit $f'(x)=0$ pour tout $x\in\mathbb{R}$ et on trouve à nouveau $f(x)=g(x)=k$ pour tout $x\in\mathbb{R}$, soit $2f(x)=f'(x)$ pour tout $x\in\mathbb{R}$ et on trouve $f(x)=g(x)=ke^{2x}$ pour tout $x\in\mathbb{R}$.
- ⋆ Dire que $(fg)'=f'g'$ revient à dire que $f'g+fg'=f'g'$. En divisant par le produit fg (il est inutile à ce stade de se préoccuper de la possibilité de diviser par 0, nous cherchons seulement formellement des conditions nécessaires) on a
$$\frac{f'(x)}{f(x)}+\frac{g'(x)}{g(x)}=\frac{f'(x)}{f(x)}\cdot\frac{g'(x)}{g(x)}$$
c'est-à-dire $\frac{f'(x)}{f(x)}=\frac{\frac{g'(x)}{g(x)}}{1-\frac{g'(x)}{g(x)}}$, soit encore
$$[\ln(f(x))]'=\frac{g'(x)}{g'(x)-g(x)}.$$
Si on choisit g, il suffit de poser $f=e^G$ où G est une primitive de $\frac{g'(x)}{g'(x)-g(x)}$.
Voyons quelques exemples :
  - ⋆ si on pose $g(x)=x$ alors $G(x)=\int\frac{1}{1-x}\,\mathrm{d}x=-\ln(1-x)$ et $f(x)=\frac{1}{1-x}$. Vérifions si on a bien $(fg)'=f'g'$:
$$(fg)'(x)=\left(\frac{x}{1-x}\right)'=\frac{1}{(1-x)^2}$$
$$f'(x)g'(x)=\frac{1}{(1-x)^2}$$
  - ⋆ si on pose $g(x)=x^a$ alors $G(x)=\int\frac{ax^{a-1}}{ax^{a-1}-x^a}\,\mathrm{d}x=\int\frac{ax^{a-1}}{ax^{a-1}-x^a}\,\mathrm{d}x=-a\ln(a-x)$ et $f(x)=\frac{1}{(a-x)^a}$. Vérifions si on a bien $(fg)'=f'g'$:
$$(fg)'(x)=\left(\frac{x^a}{(a-x)^a}\right)'=a^2x^{a-1}(a-x)^{-a-1}$$
$$f'(x)g'(x)=a(a-x)^{-a-1}\cdot ax^{a-1}=a^2x^{a-1}(a-x)^{-a-1}$$
  - ⋆ si on pose $g(x)=e^{ax}$ alors $G(x)=\int\frac{ae^{ax}}{ae^{ax}-e^{ax}}\,\mathrm{d}x=\frac{a}{a-1}x$ et $f(x)=e^{bx}$ où $b=a/(a-1)$. Vérifions si on a bien $(fg)'=f'g'$:
$$(fg)'(x)=(e^{bx}e^{ax})'=(e^{(a+b)x})'=(a+b)e^{(a+b)x}$$
$$f'(x)g'(x)=be^{bx}ae^{ax}=(ab)e^{(a+b)x}=(a+b)e^{(a+b)x}$$

Exercice 4.9 (Formules de réduction)
Les formules de réduction dérivent de l'application répétée de la règle d'intégration par parties.

1. Soit $n \in \mathbb{N}$ et $\alpha \in \mathbb{R}^*$, montrer que
$$\int x^n e^{\alpha x}\,\mathrm{d}x = \left(x^n - \frac{n}{\alpha}x^{n-1} + \frac{n(n-1)}{\alpha^2}x^{n-2}\cdots + (-1)^n\frac{n!}{\alpha^n}\right)\frac{e^{\alpha x}}{\alpha} + c$$
2. Soit $n \in \mathbb{N}$. Montrer que
$$\int \sin^n(x)\,\mathrm{d}x = \frac{-\sin^{n-1}(x)\cos(x)}{n} + \frac{n-1}{n}\int \sin^{n-2}(x)\,\mathrm{d}x,$$
$$\int \cos^n(x)\,\mathrm{d}x = \frac{\cos^{n-1}(x)\sin(x)}{n} + \frac{n-1}{n}\int \cos^{n-2}(x)\,\mathrm{d}x.$$
3. Soit $n \in \mathbb{N}$. Montrer que
$$\int x^n \sin(x)\,\mathrm{d}x = -x^n\cos(x) + nx^{n-1}\sin(x) - n(n-1)\int x^{n-2}\sin(x)\,\mathrm{d}x,$$
$$\int x^n \cos(x)\,\mathrm{d}x = x^n\sin(x) + nx^{n-1}\cos(x) - n(n-1)\int x^{n-2}\cos(x)\,\mathrm{d}x.$$
4. Soit $n \in \mathbb{N}^*$, $\alpha \neq -1$ et $x > 0$. Montrer que
$$\int x^\alpha \ln^n(x)\,\mathrm{d}x = \left(\ln^n(x) - \frac{n}{\alpha+1}\ln^{n-1}(x) + \frac{n(n-1)}{(\alpha+1)^2}\ln^{n-2}(x)\cdots + (-1)^n\frac{n!}{(\alpha+1)^n}\right)\frac{x^{\alpha+1}}{\alpha+1} + c.$$

Correction

1. On pose $I_n = \int x^n e^{\alpha x}\,\mathrm{d}x$. En intégrant par parties ($f(x) = x^n$ et $g'(x) = e^{\alpha x}$) on trouve
$$I_n = x^n\frac{e^{\alpha x}}{\alpha} - \frac{n}{\alpha}I_{n-1} = x^n\frac{e^{\alpha x}}{\alpha} - \frac{n}{\alpha}\left(x^{n-1}\frac{e^{\alpha x}}{\alpha} - \frac{n-1}{\alpha}I_{n-2}\right) = \cdots = \left(x^n - \frac{n}{\alpha}x^{n-1} + \frac{n(n-1)}{\alpha^2}x^{n-2}\cdots + (-1)^n\frac{n!}{\alpha^n}\right)\frac{e^{\alpha x}}{\alpha} + c$$
2. On pose $I_n = \int \sin^n(x)\,\mathrm{d}x$. En intégrant par parties ($f(x) = \sin^{n-1}(x)$ et $g'(x) = \sin(x)$) on trouve
$$I_n = -\sin^{n-1}(x)\cos(x) + (n-1)I_{n-2} - (n-1)I_n = \frac{-\sin^{n-1}(x)\cos(x)}{n} + \frac{n-1}{n}I_{n-2}$$
De la même manière, on pose $I_n = \int \cos^n(x)\,\mathrm{d}x$. En intégrant par parties ($f(x) = \cos^{n-1}(x)$ et $g'(x) = \cos(x)$) on trouve
$$I_n = \frac{\cos^{n-1}(x)\sin(x)}{n} + \frac{n-1}{n}I_{n-2}$$
3. On pose $I_n = \int x^n\sin(x)\,\mathrm{d}x$ et $J_n = \int x^n\cos(x)\,\mathrm{d}x$. En intégrant par parties ($f(x) = x^n$ et $g'(x) = \sin(x)$ dans la première intégrale et $f(x) = x^n$ et $g'(x) = \cos(x)$ dans la deuxième intégrale) on trouve
$$I_n = -x^n\cos(x) + nJ_{n-1} \qquad J_n = x^n\sin(x) - nI_{n-1}$$
Par conséquence
$$I_n = -x^n\cos(x) + n\left(x^{n-1}\sin(x) - (n-1)I_{n-2}\right) = -x^n\cos(x) + nx^{n-1}\sin(x) - n(n-1)I_{n-2}$$
$$J_n = x^n\sin(x) - n\left(-x^{n-1}\cos(x) + (n-1)J_{n-2}\right) = x^n\sin(x) + nx^{n-1}\cos(x) - n(n-1)J_{n-2}$$
4. On pose $I_n = \int x^\alpha\ln^n(x)\,\mathrm{d}x$. En intégrant par parties ($f(x) = \ln^n(x)$ et $g'(x) = x^\alpha$) on trouve
$$I_n = \frac{x^{\alpha+1}}{\alpha+1}\ln^n(x) - \frac{n}{\alpha+1}I_{n-1} = \cdots = \left(\ln^n(x) - \frac{n}{\alpha+1}\ln^{n-1}(x) + \frac{n(n-1)}{(\alpha+1)^2}\ln^{n-2}(x)\cdots + (-1)^n\frac{n!}{(\alpha+1)^n}\right)\frac{x^{\alpha+1}}{\alpha+1} + c.$$

Intégrales : aires, déplacements, vitesses, accélérations

Exercice 4.10 (Vitesse et accélération)
Soit $V > 0$ une constante. Une voiture roule à une vitesse de $v(t) = Vt(1-t)$ km h^{-1} durant l'intervalle de temps $0 \le t \le 1$ h. Quelle a été sa vitesse maximale ? Que vaut l'accélération instantanée ? Quelle distance a-t-elle parcouru ?

Correction

- ⋆ Vitesse maximale : $v(t) = Vt(1-t) = V(t-t^2)$, $v'(t) = V(1-2t)$, $v'(t) = 0$ ssi $t = \frac{1}{2}$ et $v\left(\frac{1}{2}\right) = \frac{V}{4}$.
- ⋆ Accélération instantanée : $a(t) = v'(t = V(1-2t)$ donc $a > 0$ si $t < 1/2$ et $a < 0$ si $t > 1/2$
- ⋆ Distance parcourue : $v(t) = x'(t)$ donc $x_{\text{parcourue}} = \displaystyle\int_0^1 v(t)\,\mathrm{d}t = \frac{V}{6}$.

Exercice 4.11

Calculer

$$\mathscr{A} = \int_{-1}^{1} \frac{1}{1+x^2}\,\mathrm{d}x, \qquad \mathscr{B} = \int_{-1/\sqrt{3}}^{1/\sqrt{3}} \frac{1}{1+3x^2}\,\mathrm{d}x, \qquad \mathscr{C} = \int_{-1}^{1} \frac{2x-5}{x^2-5x+6}\,\mathrm{d}x.$$

Correction

$$\mathscr{A} = [\arctan(x)]_{-1}^{1} = \arctan(1) - \arctan(-1) = \frac{\pi}{4} - \frac{-\pi}{4} = \frac{\pi}{2},$$

$$\mathscr{B} = \frac{1}{\sqrt{3}} \int_{-1}^{1} \frac{1}{1+t^2}\,\mathrm{d}t = \frac{\mathscr{A}}{\sqrt{3}} = \frac{\pi}{2\sqrt{3}},$$

$$\mathscr{C} = \int_{-1}^{1} \frac{2x-5}{x^2-5x+6}\,\mathrm{d}x = \left[\ln|x^2-5x+6|\right]_{-1}^{1} = \ln(2) - \ln(12) = \ln(2) - \ln(3) - 2\ln(2) = -\ln(2) - \ln(3).$$

Exercice 4.12 (Calcul de l'aire)

Calculer l'aire comprise entre le graphe de la fonction $f(x)$ et le graphe de la fonction $g(x)$:

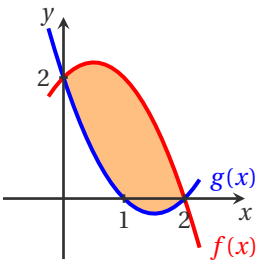
a) $f(x) = -x^2 + x + 2$ et $g(x) = x^2 - 3x + 2$ b) $f(x) = \frac{x^3}{4}$ et $g(x) = x^2 - x$

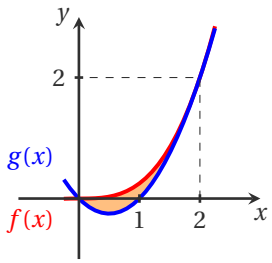
Correction

a) Comme $f(x) = g(x)$ ssi $x \in \{0,2\}$ et $f(x) \geq g(x)$ pour $x \in [0,2]$, l'aire comprise entre le graphe de la fonction $f(x)$ et le graphe de la fonction $g(x)$ est $\displaystyle\int_0^2 (f(x) - g(x))\,\mathrm{d}x = \int_0^2 -2x^2 + 4x\,\mathrm{d}x = \left[-2\frac{x^3}{3} + 4\frac{x^2}{2}\right]_0^2 = \frac{8}{3}$.

b) Comme $f(x) = g(x)$ ssi $x \in \{0,2\}$ et $f(x) \geq g(x)$ pour $x \in [0,2]$, l'aire comprise entre le graphe de la fonction $f(x)$ et le graphe de la fonction $g(x)$ est $\displaystyle\int_0^2 (f(x) - g(x))\,\mathrm{d}x = \int_0^2 \frac{x^3}{4} - x^2 + x\,\mathrm{d}x = \left[\frac{x^4}{16} - \frac{x^3}{3} + \frac{x^2}{2}\right]_0^2 = \frac{1}{3}$.

$f(x) = -x^2 + x + 2 = -(x+1)(x-2)$ et $g(x) = x^2 - 3x + 2 = (x-1)(x-2)$ $f(x) = \frac{x^3}{4}$ et $g(x) = x^2 - x = x(x-1)$





Exercice 4.13 (Calcul de l'aire)

Calculer l'aire de A et de B ainsi définis :

$$A = \left\{ (x,y) \in \mathbb{R}^2 \;\middle|\; 0 \leq x \leq 2\pi, \frac{1}{8} \leq y \leq \cos^3(x) \right\}, \qquad B = \left\{ (x,y) \in \mathbb{R}^2 \;\middle|\; -\pi \leq x \leq \pi, \frac{1}{8} \leq y \leq \cos^3(x) \right\}.$$

Correction
Remarquons d'abord que

$$\begin{aligned}\int \cos^3(x)\,\mathrm{d}x &\overset{\cos^2(x)=1-\sin^2(x)}{=} \int (1-\sin^2(x))\cos(x)\,\mathrm{d}x = \\ &= \int \cos(x) - \cos(x)\sin^2(x)\,\mathrm{d}x = \sin(x) - \int \cos x\,\sin^2(x)\,\mathrm{d}x \overset{\substack{t=\sin(x)\\ \mathrm{d}t=\cos(x)\,\mathrm{d}x}}{=} \\ &= \sin(x) - \int t^2\,\mathrm{d}t = \sin(x) - \frac{t^3}{3} + k = \sin(x) - \frac{\sin^3(x)}{3} + k\,, \quad k\in\mathbb{R}.\end{aligned}$$

Comme $\cos^3(x) = \frac{1}{8}$ ssi $x = \frac{\pi}{3}$ ou $x = \frac{5\pi}{3}$ alors

$$\text{Aire}\,(A) = 2\left(\int_0^{\frac{\pi}{3}} \cos^3(x)\,\mathrm{d}x - \frac{\frac{\pi}{3}-0}{8}\right) = 2\left(\frac{\sqrt{3}}{2} - \frac{\sqrt{3}}{8} - \frac{\pi}{24}\right) = \frac{9\sqrt{3}-\pi}{12}$$

et

$$\text{Aire}\,(B) = \int_{-\frac{\pi}{3}}^{\frac{\pi}{3}} \cos^3(x)\,\mathrm{d}x - 2\cdot\frac{\frac{\pi}{3}-0}{8} = 2\left(\frac{\sqrt{3}}{2} - \frac{\sqrt{3}}{8} - \frac{\pi}{24}\right) = \frac{9\sqrt{3}-\pi}{12}.$$

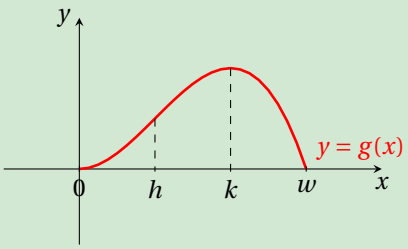
Exercice 4.14
La valeur moyenne de la fonction $f(x) = x^3$ sur l'intervalle $[0;k]$ est 9. Calculer k.

Correction
valeur moyenne de $f \overset{\text{def}}{=} \frac{1}{k}\int_0^k x^3\,\mathrm{d}x = 9 \quad\Longrightarrow\quad \frac{1}{k}\frac{k^4}{4} = 9 \quad\Longrightarrow\quad k = \sqrt[3]{36}$.

Exercice 4.15
Dans la figure ci-contre on a tracé le graphe de la fonction $g\colon [0;w] \to \mathbb{R}$ définie par

$$g(x) = \int_0^x f(t)\,\mathrm{d}t$$

avec $f\colon [0;w] \to \mathbb{R}$ une fonction continue et dérivable. Le graphe de g a tangente horizontale en $x = 0$ et présente un changement de concavité en $x = h$ et un maximum en $x = k$.



1. Calculer $f(0)$ et $f(k)$.
2. Tracer un graphe plausible de f et montrer qu'elle admet un maximum.
3. Dorénavant on suppose que g est une fonction polynomiale de degré 3.
   3.1. Montrer que $h = w/3$ et $k = 2h$.
   3.2. Pour $w = 3$ et $g(1) = 2/3$ trouver l'expression de g.

Correction

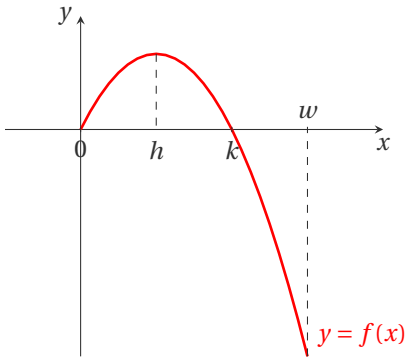
1. $g'(x) = f(x)$ pour tout $x \in [0;w]$. Puisque $x = 0$ et $x = k$ sont des points à tangente horizontale pour le graphe de g, alors $g'(0) = g'(k) = 0$ donc $f(0) = f(k) = 0$.

2. f est continue par hypothèse. D'après le théorème de WEIERSTRASS elle admet un maximum et un minimum sur $[0;w]$.

   On a vu au point précédent que $f(0) = f(k) = 0$ et que $g'(x) = f(x)$ pour tout $x \in [0;w]$. g est croissante ($g'(x) > 0$) sur $[0;k]$ et décroissante ($g'(x) < 0$) sur $[k;w]$, donc f est positive sur $[0;h]$ et négative sur $[h;w]$.

   De plus, $g''(x) = f'(x)$ pour tout $x \in [0;w]$. g est convexe ($g''(x) > 0$) sur $[0;h]$ et concave ($g''(x) < 0$) sur $[h;w]$, donc f est croissante sur $[0;h]$ et décroissante sur $[h;w]$. $x = h$ est un maximum absolu pour f et $x = w$ un minimum absolu.

   Un graphe plausible de f est donc le suivant :



3. g est une fonction polynomiale de degré 3, $x = w$ est un zéro simple et $x = 0$ est un zéro double (car $g'(0) = 0$), donc $g(x) = a(x-w)(x-0)^2 = ax^2(x-w)$ avec $a \in \mathbb{R}^*$ un paramètre.

   3.1. On a $f(x) = g'(x) = ax(3x-2w)$: il s'agit d'une parabole. Comme $f(0) = f(k) = 0$ alors $k = 2w/3$. De plus, le sommet de la parabole se trouve en $x = k/2$ et $f'(k/2) = 0$. Comme $x = h$ est le maximum de f, alors $h = k/2 = w/3$.

   3.2. Si $w = 3$ alors $g(x) = ax^2(x-3)$ et la condition $g(1) = 2/3$ implique $a = -1/3$. On obtient ainsi $g(x) = -\frac{1}{3}x^3 + x^2$.

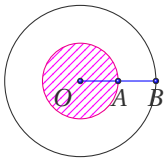
Exercice 4.16 (Probabilité géométrique)

1. On sélectionne un point au hasard sur une cible *circulaire*. Quelle est la probabilité que le point choisi soit plus prêt du centre que de la circonférence de la cible ?
2. On sélectionne un point au hasard sur une cible *carrée*. Quelle est la probabilité que le point sélectionné soit plus prêt du centre du carré que d'un de ses côtés ?

Source : http://www.thedudeminds.net

Correction

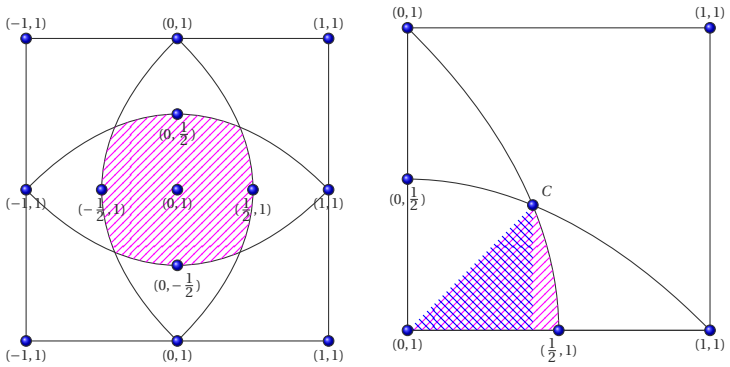
1. Il semble assez évident de délimiter correctement de manière intuitive les zones par deux disques concentriques.



   Attention néanmoins à ne pas répondre que la probabilité de choisir un point dans la zone hachurée est 1/2 (parce que le rayon du disque hachuré correspond à la moitié du rayon du grand disque). Or, si le petit disque possède un rayon de r et le grand $2r$, on a

$$P = \frac{\text{Aire du petit disque hachuré}}{\text{Aire du grand disque}} = \frac{\pi r^2}{\pi(2r)^2} = \frac{1}{4}.$$

2. L'ensemble de points équidistants d'un point et d'une droite est une parabole. La région à considérer est donc délimitée par quatre paraboles qui ont pour foyer le centre du carré et comme droites directrices les droites qui supportent les côtés du carré. On s'affaire donc à trouver l'aire de cette région hachurée. On place d'abord le tout dans un repère cartésien. Les sommets du carré sont $(1,1)$, $(-1,1)$, $(-1,-1)$ et $(1,-1)$ (cf. figure à gauche). En vertu des symétries de la figure, il nous est possible de nous concentrer seulement sur la partie située dans le premier quadrant. Qui plus est, il est possible de ne s'attarder qu'à la moitié de cette dernière région (cf. figure à droite).



Ce «croissant de parabole» correspond à la moitié de la région à considérer dans le premier quadrant. On note au passage que l'aire du carré dans ce premier quadrant est 1. Le «croissant» est à son tour divisé en deux parties : la zone de forme triangulaire $\mathscr{T}$ et la zone $\mathscr{A}$. On cherche l'aire de ces zones. Pour y arriver, on aura besoin des coordonnées de C. L'équation de la parabole qui nous intéresse est $y^2 = 1-2x$. On ne s'intéressera qu'à la branche située au dessus de l'axe des abscisses. Ainsi $y = \sqrt{1-2x}$. On cherche ensuite les coordonnées de C, le point d'intersection entre la courbe d'équation $y = \sqrt{1-2x}$ et la droite d'équation $y = x$ et on obtient $x = -1+\sqrt{2}$. Il nous est donc déjà possible de trouver l'aire du triangle, que l'on a identifié comme $\mathscr{T} = \frac{3-2\sqrt{2}}{2}$. Il reste à trouver l'aire de la région sous la courbe :

$$\mathscr{A} = \int_{-1+\sqrt{2}}^{\frac{1}{2}} \sqrt{1-2x}\,\mathrm{d}x = -\frac{1}{2}\int_{3-2\sqrt{2}}^{0} \sqrt{t}\,\mathrm{d}t = \frac{1}{3}\sqrt{\left(3-2\sqrt{2}\right)^3} = \frac{(-1+\sqrt{2})^3}{3}.$$

L'aire totale est donc

$$\mathscr{T} + \mathscr{A} = \frac{-5+4\sqrt{2}}{6}.$$

Comme le carré du premier quadrant à une aire de 1, il ne reste qu'à doubler ce résultat afin d'obtenir la probabilité recherchée

$$P = \frac{-5+4\sqrt{2}}{3}$$

ce qui correspond à un peu moins de 22%.

4.3.2. Calcul approché d'intégrales

Exercice 4.17
Estimer $\int_0^{5/2} f(x)\,\mathrm{d}x$ 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 $h = \frac{b-a}{m} = \frac{1}{2}$.

```
h=1/2;
x=[0:h:5/2]
y=[3/2 2 2 1.6364 1.2500 0.9565]
Gauche=h*sum( y(1:end-1) )
Droite=h*sum( y(2:end) )
Trapeze=0.5*h*( y(1)+2*sum(y(2:end-1))+y(end) )
```

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 4.18
Étant donnée l'égalité

$$\pi = 4\left(\int_0^{+\infty} e^{-x^2}\,\mathrm{d}x\right)^2 = 4\left(\int_0^{10} e^{-x^2}\,\mathrm{d}x + \epsilon\right)^2,$$

avec $0 < \epsilon < 10^{-44}$, utiliser la méthode des trapèzes composite à 10 intervalles pour estimer la valeur de π.

Correction
La méthode des trapèzes composite à m intervalles pour calculer l'intégrale d'une fonction f sur l'intervalle $[a,b]$ s'écrit

$$\int_a^b f(t)dt \simeq h\left(\frac{1}{2}f(a) + \sum_{i=1}^{m-1} f(a+ih) + \frac{1}{2}f(b)\right) \qquad \text{avec } h = \frac{b-a}{m}.$$

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 $\pi \approx 4I^2 = 3.1422$.

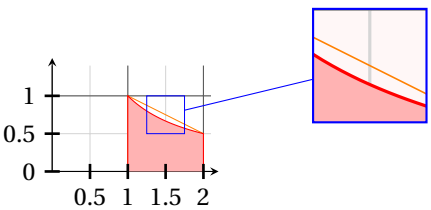
Exercice 4.19
On considère l'intégrale

$$I = \int_1^2 \frac{1}{x}\,\mathrm{d}x.$$

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 $I = \Big[\ln(x)\Big]_{x=1}^{x=2} = \ln(2)$.
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
$$\int_a^b f(t)dt \simeq h\left(\frac{1}{2}f(a) + \sum_{i=1}^{m-1} f(a+ih) + \frac{1}{2}f(b)\right) \qquad \text{avec } h = \frac{b-a}{m}.$$
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 précédant l'aire sous les trapèzes sera supérieure à l'aire exacte.

Exercice 4.20 (Interpolation, Quadrature et EDO)

1. Soit f une fonction de classe $\mathscr{C}^1([-1,1])$. 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 :
$$\int_0^1 f(\tau)\,\mathrm{d}\tau \approx \int_0^1 p(\tau)\,\mathrm{d}\tau.$$
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
$$\int_a^b f(x)\,\mathrm{d}x$$
lorsque f est une fonction de classe $\mathscr{C}^1([2a-b,b])$.
Remarque : $[2a-b,b] = [a-(b-a), a+(b-a)]$
4. Considérons le problème de CAUCHY : trouver y : $[t_0,T] \subset \mathbb{R} \to \mathbb{R}$ 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 $h = \dfrac{T-t_0}{N}$ avec $t_n = t_0 + nh$ pour $n = 0,\dots,N$.
Utiliser la formule obtenue au point 3 pour approcher l'intégrale
$$\int_{t_n}^{t_{n+1}} \varphi(t, y(t))\,\mathrm{d}t.$$
En déduire un schéma à deux pas implicite pour l'approximation de la solution du problème de CAUCHY.

Correction

1. On cherche les coefficients α, β et γ 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
$$\int_0^1 f(\tau)\,\mathrm{d}\tau \approx \int_0^1 p(\tau)\,\mathrm{d}\tau = \alpha + \frac{\beta}{2} + \frac{\gamma}{3} = \frac{-f(-1)+8f(0)+5f(1)}{12}.$$
3. Soit $x = m\tau + q$, alors
$$\int_a^b f(x)\,\mathrm{d}x = m\int_0^1 f(m\tau+q)\,\mathrm{d}\tau \quad \text{avec} \quad \begin{cases} a = q, \\ b = m+q, \end{cases} \quad \textit{i.e.} \quad \begin{cases} m = b-a, \\ q = a, \end{cases}$$
d'où le changement de variable $x = (b-a)\tau + a$. On en déduit la formule de quadrature
$$\int_a^b f(x)\,\mathrm{d}x = (b-a)\int_0^1 f\left((b-a)\tau + a\right)\,\mathrm{d}\tau \approx (b-a)\frac{-f(2a-b)+8f(a)+5f(b)}{12}.$$
4. On pose $a = t_n$ et $b = t_{n+1}$ d'où la formule de quadrature
$$\int_{t_n}^{t_{n+1}} f(t)\,\mathrm{d}t \approx (t_{n+1}-t_n)\frac{-f(2t_n-t_{n+1})+8f(t_n)+5f(t_{n+1})}{12} = h\frac{-f(t_{n-1})+8f(t_n)+5f(t_{n+1})}{12}.$$

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

$$y(t_{n+1})=y(t_n)+\int_{t_n}^{t_{n+1}}\varphi(t,y(t))\,\mathrm{d}t\approx h\frac{-\varphi(t_{n-1},y(t_{n-1}))+8\varphi(t_n,y(t_n))+5\varphi(t_{n+1},y(t_{n+1}))}{12}.$$

Si on note u_n une approximation de la solution y au temps t_n, on obtient le schéma à deux pas implicite suivant :

$$\begin{cases} u_0=y(t_0)=y_0,\\ u_1 \text{ à définir},\\ u_{n+1}=u_n+h\dfrac{-\varphi(t_{n-1},u_{n-1})+8\varphi(t_n,u_n)+5\varphi(t_{n+1},u_{n+1})}{12} & n=1,2,\dots N-1 \end{cases}$$

On peut utiliser une prédiction d'Euler explicite pour initialiser u_1 :

$$\begin{cases} u_0=y(t_0)=y_0,\\ u_1=u_0+h\varphi(t_0,u_0),\\ u_{n+1}=u_n+h\dfrac{-\varphi(t_{n-1},u_{n-1})+8\varphi(t_n,u_n)+5\varphi(t_{n+1},u_{n+1})}{12} & n=1,2,\dots N-1 \end{cases}$$

Exercice 4.21 (Interpolation, Quadrature et EDO)

1. Soit $h>0$ et $f\colon[a-h,a+h]\to\mathbb{R}$ une fonction de classe $\mathscr{C}^1([a-h,a+h])$. 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 :
$$\int_a^{a+h}f(x)\,\mathrm{d}x\approx\int_a^{a+h}p(x)\,\mathrm{d}x.$$
NB : on intègre sur $[a,a+h]$ mais on interpole en $a-h$ et a.
3. Considérons le problème de CAUCHY : trouver $y\colon[t_0,T]\subset\mathbb{R}\to\mathbb{R}$ tel que
$$\begin{cases} y'(t)=\varphi(t,y(t)), \quad \forall t\in[t_0,T],\\ y(t_0)=y_0, \end{cases}$$
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 $h=\dfrac{T-t_0}{N}$ avec $t_n=t_0+nh$ pour $n=0,\dots,N$.
Utiliser la formule obtenue au point 2 pour approcher l'intégrale
$$\int_{t_n}^{t_{n+1}}\varphi(t,y(t))\,\mathrm{d}t.$$
En déduire un schéma à deux pas explicite pour l'approximation de la solution du problème de CAUCHY.

Correction

1. $p(x)=\dfrac{f(a)-f(a-h)}{a-(a-h)}(x-a)+f(a)=\dfrac{f(a)-f(a-h)}{h}(x-a)+f(a).$
2. On en déduit la méthode de quadrature
$$\begin{aligned}\int_a^{a+h}f(x)\,\mathrm{d}x&\approx\int_a^{a+h}p(x)\,\mathrm{d}x\\ &=\frac{f(a)-f(a-h)}{h}\left[\frac{(x-a)^2}{2}\right]_a^{a+h}+f(a)\,[x]_a^{a+h}\\ &=\frac{f(a)-f(a-h)}{2h}\left((a+h-a)^2-(a-a)^2\right)+f(a)(a+h-a)\\ &=\frac{f(a)-f(a-h)}{2h}h^2+hf(a)\\ &=h\frac{3f(a)-f(a-h)}{2}.\end{aligned}$$

3. On pose $a = t_n$ et $a + h = t_{n+1}$ d'où la formule de quadrature

$$\int_{t_n}^{t_{n+1}} f(t)\,\mathrm{d}t \approx (t_{n+1} - t_n)\frac{3f(t_n) - f(2t_n - t_{n+1})}{2} = h\frac{3f(t_n) - f(t_{n-1})}{2}.$$

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

$$y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} \varphi(t, y(t))\,\mathrm{d}t \approx h\frac{3\varphi(t_n, y(t_n)) - \varphi(t_{n-1}, y(t_{n-1}))}{2}.$$

Si on note u_n une approximation de la solution y au temps t_n, on obtient le schéma à deux pas implicite suivant :

$$\begin{cases} u_0 = y(t_0) = y_0, \\ u_1 \text{ à définir}, \\ u_{n+1} = u_n + h\dfrac{3\varphi(t_{n-1}, u_{n-1}) - \varphi(t_n, u_n)}{2} \quad n = 1, 2, \dots N-1 \end{cases}$$

On peut utiliser une prédiction d'Euler explicite pour initialiser u_1 :

$$\begin{cases} u_0 = y(t_0) = y_0, \\ u_1 = u_0 + h\varphi(t_0, u_0), \\ u_{n+1} = u_n + h\dfrac{3\varphi(t_{n-1}, u_{n-1}) - \varphi(t_n, u_n)}{2} \quad n = 1, 2, \dots N-1 \end{cases}$$

CHAPITRE 5

De l'interpolation à l'approximation d'EDO

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).

5.1. EDO : généralités

Une EDO (voir par exemple [1, Ch. 8]) est une équation exprimée sous la forme d'une relation

$$F(y(t), y'(t), y''(t), \ldots, y^{(p)}(t)) = g(t)$$

- ⋆ dont les inconnues sont une **fonction** $y\colon I \subset \mathbb{R} \to \mathbb{R}$ et son **intervalle de définition** I
- ⋆ dans laquelle cohabitent à la fois la fonction inconnue y et ses dérivées y', y'', …,$y^{(p)}$ (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**.
Nous pouvons nous limiter aux équations différentielles du premier ordre, car une équation d'ordre $p > 1$ peut toujours se ramener à un système de p équations d'ordre 1.

Dans la suite nous ne considérerons que des EDO d'ordre 1 écrite sous la forme

$$y'(t) = \varphi(t, y(t)).$$

Si φ ne dépend pas explicitement de t (*i.e.* si $\varphi(t, y(t)) = \varphi(y(t))$), l'EDO est dite *autonome*.

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). [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.

Condition initiale pour une EDO d'ordre 1. 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 en un point de l'intervalle d'intégration : $y(t_0) = y_0$ impose en t_0 la valeur y_0 de la fonction inconnue. 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.

EXEMPLE
Résoudre l'équation différentielle $y'(t) = -y(t)$ signifie chercher toutes les fonctions

$$\begin{aligned} y\colon I \subset \mathbb{R} &\to \mathbb{R} \\ t &\mapsto y = f(t) \end{aligned}$$

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 solution de l'EDO, elle est même la solution générale. Si parmi toutes ces solutions, on cherche celle qui vérifie $y(0) = 0$, on trouve que c doit être nul : c'est une solution particulière.

1. Résoudre une équation c'est chercher toutes les valeurs de l'inconnue qui satisfont l'égalité. Dans les équations rencontrées jusqu'à présent, les inconnues étaient des nombres. Par exemple, résoudre l'équation $2x + 4 = 10$ signifie chercher toutes les valeurs de $x \in \mathbb{R}$ telles que $2x + 4 = 10$. Dans les équations différentielles, les inconnues sont des fonctions.

Définition 5.1 (Problème de CAUCHY)
Soit $I \subset \mathbb{R}$ un intervalle, t_0 un point de I, $\varphi\colon I \times \mathbb{R} \to \mathbb{R}$ 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 $y \in \mathscr{C}^1(I)$ telle que

$$\begin{cases} y'(t) = \varphi(t, y(t)), & \forall t \in I, \\ y(t_0) = y_0, \end{cases} \tag{5.1}$$

avec y_0 une valeur donnée appelée *donnée initiale*.

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 (*i.e.* la solution générale),
- ⋆ 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 5.2
Le problème de Cauchy (5.1) est équivalent à l'équation intégrale

$$y(t) = y_0 + \int_{t_0}^{t} \varphi(s, y(s))\, \mathrm{d}s. \tag{5.2}$$

PREUVE
En intégrant l'EDO entre t_0 et t et en considérant la donnée initiale (5.1) on obtient

$$y(t) = y_0 + \int_{t_0}^{t} \varphi(s, y(s))\, \mathrm{d}s.$$

La solution du problème de Cauchy est donc de classe $\mathscr{C}^1(I)$ sur I et satisfait l'équation intégrale (5.2).
Inversement, si y est définie par (5.2), 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 $\mathscr{C}^1(I)$ et satisfait l'équation différentielle $y'(t) = \varphi(t, y(t))$.
Ainsi, si φ est continue, le problème de Cauchy (5.1) est équivalent à l'équation intégrale (5.2).

Nous verrons plus loin comment tirer parti de cette équivalence pour les méthodes numériques.

5.1.1. Existence et unicité

Considérons un exemple de problème de Cauchy :

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 $y\colon t \in \mathbb{R} \mapsto y(t) \in \mathbb{R}$ 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

$$y(0) = (\alpha + 1/3)e^{0} + 0 - 1/3 = \alpha, \qquad y'(t) = -3(\alpha + 1/3)e^{-3t} + 1 = -3(\alpha + 1/3)e^{-3t} + 1 - 3t + 3t = -3y(t) + 3t.$$

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 $y\colon t \in \mathbb{R}^+ \mapsto y(t) \in \mathbb{R}$ 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 $y_{2,3}(t) = \pm\sqrt{8t^3/27}$, 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 $y\colon\ t \in \mathbb{R}^+ \mapsto y(t) \in \mathbb{R}$ qui satisfait

$$\begin{cases} y'(t) = |y(t)|^\alpha, & \forall\, t > 0, \\ y(0) = 0. \end{cases}$$

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 $y\colon\ t \in \mathbb{R}^+ \mapsto y(t) \in \mathbb{R}$ 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 $y(t) = \frac{1}{\sqrt{1-2t}}$ 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 $\lim\limits_{t\to(1/2)^-} y(t) = +\infty$) : 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 ne considérerons que des problèmes de CAUCHY admettant une et une seule solution sur l'intervalle indiqué.

5.2. Calcul analytique des solutions de quelques types d'EDO d'ordre 1

5.2.1. 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

$$\underbrace{\int f(y(x))\,y'(x)\,\mathrm{d}x}_{=\int f(u)\,\mathrm{d}u = F(u)} = \underbrace{\int g(x)\,\mathrm{d}x}_{G(x)+C}$$

donc

$$F(y(x)) = G(x) + C \qquad \text{où } C \in \mathbb{R},$$

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 $y'(x) = \mathrm{d}y/\mathrm{d}x$, on peut écrire l'équation $f(y(x))y'(x) = g(x)$ sous la forme

$$f(y)\,\mathrm{d}y = g(x)\,\mathrm{d}x,$$

puis intégrer formellement les deux membres

$$\int f(y)\,\mathrm{d}y = \int g(x)\,\mathrm{d}x,$$

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 :

$$\frac{y'(x)}{y(x)} = x \quad\Longrightarrow\quad \frac{\frac{\mathrm{d}y}{\mathrm{d}x}}{y(x)} = x \quad\Longrightarrow\quad \frac{\mathrm{d}y}{y} = x\,\mathrm{d}x \quad\Longrightarrow\quad \int \frac{1}{y}\,\mathrm{d}y = \int x\,\mathrm{d}x \quad\Longrightarrow\quad \ln|y| = \frac{x^2}{2} + C \text{ avec } C \in \mathbb{R}.$$

  Ainsi, toute solution non nulle est de la forme

$$y(x) = De^{x^2/2} \qquad \text{avec } D \in \mathbb{R}^*.$$

5.2.2. 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 $x \in \mathscr{D}_b \cap \mathscr{D}_g \cap \{x \in \mathscr{D}_a \mid a(x) \neq 0\}$, 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) :

$$y(x) = \underbrace{Ce^{-A(x)}}_{y_H(x)} + \underbrace{B(x)e^{-A(x)}}_{y_P(x)}$$

avec

- $A(x)$ une primitive de $\dfrac{b(x)}{a(x)}$,
- $B(x)$ une primitive de $\dfrac{g(x)}{a(x)}e^{A(x)}$.

PREUVE
Pour vérifier que c'est bien une solution il suffit de dériver :

$$\begin{aligned} y'(x) &= CA'(x)e^{-A(x)} - B(x)A'(x)e^{-A(x)} + B'(x)e^{-A(x)} \\ &= -A'(x)\left(Ce^{-A(x)} + B(x)e^{-A(x)}\right) + B'(x)e^{-A(x)} \\ &= -A'(x)y(x) + B(x)e^{-A(x)} \\ &= -\frac{b(x)}{a(x)}y(x) + \frac{g(x)}{a(x)}e^{A(x)}e^{-A(x)} = -\frac{b(x)}{a(x)}y(x) + \frac{g(x)}{a(x)} \end{aligned}$$

donc $a(x)y'(x) = -b(x)y(x) + g(x)$.

Remarque
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 :

$$\frac{y'(x)}{y(x)}=-\frac{b(x)}{a(x)} \quad\Longrightarrow\quad \int\frac{1}{y}\,\mathrm{d}y=-\int-\frac{b(x)}{a(x)}\,\mathrm{d}x \quad\Longrightarrow\quad \ln|y|=-\int\frac{b(x)}{a(x)}\,\mathrm{d}x.$$

  Ainsi, toute solution non nulle de l'équation homogène associée est de la forme

$$y_H(x)=Ce^{-A(x)} \qquad \text{où } A(x)=\int\frac{b(u)}{a(u)}\,\mathrm{d}u$$

  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

$$a(x)=1, \qquad b(x)=-1, \qquad g(x)=x.$$

Pour $x\in\mathbb{R}$ on a

- ⋆ $A(x)=\int -1\,\mathrm{d}x=-x$,
- ⋆ $B(x)=\int xe^{-x}\,\mathrm{d}x=-(1+x)e^{-x}$,

donc

$$y(x)=\left(C-(1+x)e^{-x}\right)e^{x}=Ce^{x}-(1+x).$$

EXEMPLE (LOI DE NEWTON ☕)
Considérons une tasse de café à la température de 75°C dans une salle à 25°C. Après 5 minutes le café est à 50°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
   - ⋆ $A(t)=\int-\gamma\,\mathrm{d}t=-\gamma t$,
   - ⋆ $B(t)=\int-25\gamma e^{A(t)}\,\mathrm{d}t=25\int-\gamma e^{-\gamma t}\,\mathrm{d}t=25e^{-\gamma t}$.

   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 :

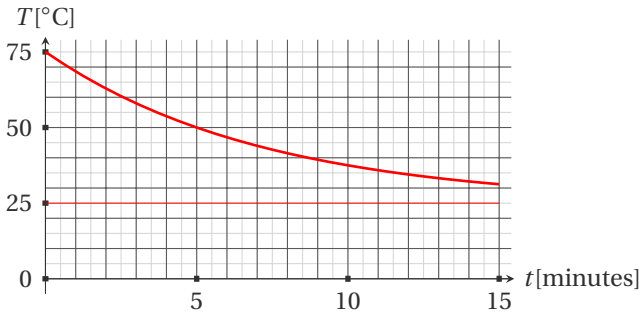
$$75=T(0)=25+De^{\gamma\cdot 0} \qquad\Longrightarrow\qquad D=50 \qquad\Longrightarrow\qquad T(t)=25+50e^{\gamma t}.$$

3. Il ne reste qu'à établir la valeur numérique de la constante de refroidissement γ grâce à l'«indice» :

$$50 = T(5) = 25 + 50e^{\gamma t} \qquad \Longrightarrow \qquad \gamma = -\frac{\ln(2)}{5} \qquad \Longrightarrow \qquad T(t) = 25 + 50e^{-\frac{\ln(2)}{5}t}$$

On peut donc conclure que la température du café évolue selon la fonction

$$T(t) = 25 + 50e^{-\frac{\ln(2)}{5}t}.$$



5.2.3. EDO d'ordre 1 de Bernoulli

Elles sont du premier ordre et de la forme

$$u(x)y'(x) + v(x)y(x) = w(x)(y(x))^{\alpha}, \qquad \alpha \in \mathbb{R} \setminus \{0;1\}$$

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 : [2]

$$\underbrace{u(x)}_{a(x)}z'(x) + \underbrace{(1-\alpha)v(x)}_{b(x)}z(x) = \underbrace{(1-\alpha)w(x)}_{g(x)}.$$

Par conséquent, pour $x \in \mathscr{D}_v \cap \mathscr{D}_w \cap \{x \in \mathscr{D}_u \mid u(x) \neq 0\}$, 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 $(1-\alpha)\dfrac{v(x)}{u(x)}$,
- $B(x)$ une primitive de $(1-\alpha)\dfrac{w(x)}{u(x)}e^{A(x)}$.

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 $\mathbb{R}$.

- $A(x) = (1-\alpha)\displaystyle\int \frac{v(x)}{u(x)}\,\mathrm{d}x = -2\int \frac{1/2}{1}\,\mathrm{d}x = -x$,
- $B(x) = (1-\alpha)\displaystyle\int \frac{w(x)}{u(x)}e^{A(x)}\,\mathrm{d}x = -2\int \frac{(x-1)/2}{1}e^{-x}\,\mathrm{d}x = \int (1-x)e^{-x}\,\mathrm{d}x = -(1-x)e^{-x} - \int e^{-x}\,\mathrm{d}x = xe^{-x}$,
- $z(x) = (C + B(x))\,e^{-A(x)} = (C + xe^{-x})e^{x} = Ce^{x} + x$,

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 $x + Ce^{x} > 0$.

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 $y' = (1-\alpha)z'y^{\alpha}$.

5.2.4. Équations différentielles linéaires homogène d'ordre 2

Considérons l'équation différentielle

$$x''(t) + Ax'(t) + Bx(t) = 0.$$

Si on pose $y_1(t) \overset{\text{def}}{=} x(t)$ et $y_2(t) \overset{\text{def}}{=} x'(t)$, elle est équivalente au système

$$\begin{cases} y_1'(t) = x'(t) = y_2(t), \\ y_2'(t) = x''(t) = -Ax'(t) - Bx(t) = -By_1(t) - Ay_2(t), \end{cases} \quad \text{soit encore } \mathbf{y}'(t) = \begin{pmatrix} 0 & 1 \\ -B & -A \end{pmatrix} \mathbf{y}(t)$$

On peut prouver que la solution générale de ce système s'écrit

$$\mathbf{y}(t) = e^{\lambda_1 t}\mathbf{v}^1 + e^{\lambda_2 t}\mathbf{v}^2$$

où λ_i est une valeur propre de

$$\mathbb{A} \overset{\text{def}}{=} \begin{pmatrix} 0 & 1 \\ -B & -A \end{pmatrix}$$

et $\mathbf{v}^i$ la vecteur propre associée.
On trouve ainsi que λ_i est une solution de $\lambda^2 + A\lambda + B = 0$.

EXEMPLE
Calculer la solution générale $t \mapsto x(t)$ de l'équation différentielle

$$x''(t) + 2x'(t) - 3x(t) = 0.$$

La matrice $\mathbb{A}$ est $\left(\begin{smallmatrix} 0 & 1 \\ 3 & -2 \end{smallmatrix}\right)$. Le polynôme caractéristique est donc

$$p(\lambda) = \det \begin{pmatrix} 0-\lambda & 1 \\ 3 & -2-\lambda \end{pmatrix} = (0-\lambda)(-2-\lambda) - 3 = \lambda^2 + 2\lambda - 3 = (\lambda - 1)(\lambda + 3)$$

donc on a les deux valeurs propres $\lambda_1 = 1$ et $\lambda_2 = -3$.
Calculons maintenant les vecteurs propres associés : $\mathbf{v}^i$ est un vecteur propre associé à la valeur propre λ_i si

$$\begin{pmatrix} 0-\lambda_i & 1 \\ 3 & -2-\lambda_i \end{pmatrix} \mathbf{v}^i = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

- ⋆ Pour $\lambda_1 = 1$ on doit résoudre

$$\begin{pmatrix} -1 & 1 \\ 3 & -3 \end{pmatrix} \begin{pmatrix} v_1^1 \\ v_2^1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

  donc $\mathbf{v}^1 = (\kappa_1, \kappa_1)^T$.
- ⋆ Pour $\lambda_2 = -3$ on doit résoudre

$$\begin{pmatrix} 3 & 1 \\ 3 & 1 \end{pmatrix} \begin{pmatrix} v_1^2 \\ v_2^2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

  donc $\mathbf{v}^2 = (\kappa_2, -3\kappa_2)^T$.

On conclut que

$$\mathbf{y}(t) = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}(t) = e^{\lambda_1 t}\mathbf{v}^1 + e^{\lambda_2 t}\mathbf{v}^2 = \begin{pmatrix} \kappa_1 \\ \kappa_1 \end{pmatrix} e^t + \begin{pmatrix} \kappa_2 \\ -3\kappa_2 \end{pmatrix} e^{-3t} \quad \text{soit encore} \quad \begin{cases} y_1(t) = \kappa_1 e^t + \kappa_2 e^{-3t}, \\ y_2(t) = \kappa_1 e^t - 3\kappa_2 e^{-3t}, \end{cases}$$

et enfin

$$x(t) = \kappa_1 e^t + \kappa_2 e^{-3t}.$$

Vérifions notre résultat :

$$\begin{aligned} x(t) &= \kappa_1 e^t + \kappa_2 e^{-3t}, \\ x'(t) &= \kappa_1 e^t - 3\kappa_2 e^{-3t}, \\ x''(t) &= \kappa_1 e^t + 9\kappa_2 e^{-3t}, \\ x''(t) + 2x'(t) - 3x(t) &= \kappa_1 e^t + 9\kappa_2 e^{-3t} + 2\kappa_1 e^t - 6\kappa_2 e^{-3t} - 3\kappa_1 e^t - 3\kappa_2 e^{-3t} = 0 \quad \forall \kappa_1, \kappa_2 \in \mathbb{R}. \end{aligned}$$

5.3. Quelques schémas numériques

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.

Considérons le problème de CAUCHY (5.1) :

trouver une fonction $y\colon I \subset \mathbb{R} \to \mathbb{R}$ définie sur un intervalle I telle que

$$\begin{cases} y'(t) = \varphi(t, y(t)), & \forall t \in I =]t_0, T[, \\ y(t_0) = y_0, \end{cases}$$

avec y_0 une valeur donnée et supposons que l'on ait montré l'existence et l'unicité d'une solution y pour $t \in I$.
On subdivise l'intervalle $I = [t_0; T]$, avec $T < +\infty$, en N intervalles $[t_n; t_{n+1}]$ de largeur $h = \frac{T-t_0}{N}$ avec $t_n = t_0 + nh$ pour $n = 0, \dots, N$. La longueur h est appelé le *pas de discrétisation*.
Pour chaque nœud t_n, on note $y_n = y(t_n)$ et on cherche la valeur inconnue u_n qui approche la valeur exacte y_n ; l'ensemble des valeurs $\{y_0, y_1, \dots, y_N\}$ représente la solution exacte discrète tandis que l'ensemble des valeurs $\{u_0 = y_0, u_1, \dots, u_N\}$ représente la solution numérique. Cette solution approchée sera obtenue en construisant une suite définie par récurrence. Les schémas qu'on va construire permettent de calculer (explicitement ou implicitement) u_{n+1} à partir de $u_n, u_{n-1}, \dots, u_{n-k}$ et il est donc possible de calculer successivement u_1, u_2,..., en partant de u_0 par une formule de récurrence de la forme

$$\begin{cases} u_0 = y_0, \\ u_{n+1} = \Phi(u_{n+1}, u_n, u_{n-1}, \dots, u_{n-k}), & \forall n \in \mathbb{N}. \end{cases}$$

Définition 5.3 (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, $k \le n$ (ou d'une partie d'entre elles) :

$$\begin{cases} u_0 = y_0, \\ u_{n+1} = \Phi(\cancel{u_{n+1}}, u_n, u_{n-1}, \dots, u_{n-k}), & \forall n \in \mathbb{N}. \end{cases}$$

Une méthode est dite *implicite* si u_{n+1} n'est définie que par une relation implicite faisant intervenir la fonction φ :

$$\begin{cases} u_0 = y_0, \\ u_{n+1} = \Phi(u_{n+1}, u_n, u_{n-1}, \dots, u_{n-k}), & \forall n \in \mathbb{N}. \end{cases}$$

Définition 5.4 (Méthodes à un pas et méthodes multi-pas)
Une méthode numérique pour l'approximation du problème de CAUCHY (5.1) est dite *à un pas* si pour tout $n \in \mathbb{N}$, u_{n+1} ne dépend que de u_n et éventuellement de lui-même :

$$\begin{cases} u_0 = y_0, \\ u_{n+1} = \Phi(u_{n+1}, u_n), & \forall n \in \mathbb{N}. \end{cases}$$

Autrement, on dit que le schéma est une méthode *multi-pas* (ou à pas multiples) :

$$\begin{cases} u_0 = y_0, \\ u_{n+1} = \Phi(u_{n+1}, u_n, u_{n-1}, \dots, u_{n-k}), \quad \forall n \in \mathbb{N}. \end{cases}$$

5.3.1. Schémas classiques

Si nous intégrons l'EDO $y'(t) = \varphi(t, y(t))$ entre t_n et t_{n+1} nous obtenons

$$y_{n+1} - y_n = \int_{t_n}^{t_{n+1}} \varphi(t, y(t)) dt.$$

On peut construire différentes schémas selon la formule d'approximation utilisée pour approcher le membre de droite. Cette solution approchée sera obtenue en construisant une suite récurrente comme suit :

$$\begin{cases} u_0 = y_0, \\ u_{n+1} = u_n + \displaystyle\int_{t_n}^{t_{n+1}} \big(\text{un polynôme d'interpolation de } \varphi(t,u)\big)\, \mathrm{d}t. \end{cases}$$

① Si on remplace une fonction 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), on a

$$\tilde{f}(x) = f(a)$$
$$\int_a^b f(x)\, \mathrm{d}x \approx \int_a^b \tilde{f}(x)\, \mathrm{d}x = (b-a) f(a).$$

En utilisant cette formule pour approcher la fonction $t \mapsto \varphi(t, y(t))$ on a

$$\int_{t_n}^{t_{n+1}} \varphi(t, y(t)) dt \approx h\varphi(t_n, y(t_n))$$

et on obtient le ***schéma d'Euler progressif***

$$\begin{cases} u_0 = y(t_0) = y_0, \\ u_{n+1} = u_n + h\varphi(t_n, u_n) \quad n = 0, 1, 2, \dots N-1 \end{cases} \tag{5.3}$$

Il s'agit d'un schéma à 1 pas explicite car il permet d'expliciter u_{n+1} en fonction de u_n.

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 `t` le vecteur contenant la discrétisation et `u` le vecteur contenant l'approximation de y en chaque point t_n.

```
1;

function [t,u]=eulerexplicite(t0,T,y0,N,phi)
  t=linspace(t0,T,N);
  u=zeros(N,1);
  u(1)=y0;
  h=t(2)-t(1);
  for n=1:N-1
     u(n+1) = u(n)+h*phi(t(n),u(n));
  end
end

% TEST :
phi=@(t,y)[-y];
[t,u]=eulerexplicite(0,10,100,11,phi);
plot(t,u,'o-',t,100*exp(-t))
```

② Si on remplace une fonction 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), on a

$$\tilde{f}(x) = f(b)$$

$$\int_a^b f(x)\,\mathrm{d}x \approx \int_a^b \tilde{f}(x)\,\mathrm{d}x = (b-a)f(b).$$

En utilisant cette formule pour approcher la fonction $t \mapsto \varphi(t, y(t))$ on a

$$\int_{t_n}^{t_{n+1}} \varphi(t, y(t))\,dt \approx h\varphi(t_{n+1}, y(t_{n+1}))$$

et on obtient le ***schéma d'*EULER *rétrograde***

$$\begin{cases} u_0 = y(t_0) = y_0, \\ u_{n+1} - h\varphi(t_{n+1}, u_{n+1}) = u_n \quad n = 0, 1, 2, \dots N-1 \end{cases} \tag{5.4}$$

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. Pour calculer u_{n+1} il faudra utiliser un schéma pour le calcul du zéro d'une fonction quelconque (par exemple la méthode de la dichotomie).

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 renvoit `t` le vecteur contenant la discrétisation de l'intervalle $[t_0, T]$ et `u` le vecteur contenant l'approximation de y en chaque point t_n. Comme il s'agit d'un schéma implicite on calcule u_{n+1} en utilisant la fonction prédéfinie `fzero` :

- ⋆ à chaque étape n on doit résoudre une équation du type $u_{n+1} = G(u_n) + F(u_{n+1})$ d'inconnue u_{n+1} ;
- ⋆ on définit une fonction $y \mapsto G(u_n) + F(y) - y$ comme une fonction anonyme `temp=@(y)[G(u(n))+F(y)-y]` ;
- ⋆ résoudre l'équation équivaut à chercher un zéro de la fonction anonyme, on utilise alors la fonction `fzero` et comme point de départ on prendra u_n : `fzero(temp(y),u(n))`.

```
1;

function [t,u]=eulerimplicite(t0,T,y0,N,phi)
  t=linspace(t0,T,N);
  u=zeros(N,1);
  u(1)=y0;
  h=t(2)-t(1);
  for n=1:N-1
     u(n+1)=fzero(@(y) u(n)+h*phi(t(n+1),y)-y, u(n));
  end
end

% TEST :
phi=@(t,y)[-y];
[t,u]=eulerimplicite(0,10,100,11,phi);
plot(t,u,'o-',t,100*exp(-t))
```

③ Si on remplace une fonction f par une constante égale à la valeur de f au milieu de l'intervalle $[a; b]$ (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), on a

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

En utilisant cette formule pour approcher la fonction $t \mapsto \varphi(t, y(t))$ on a

$$\int_{t_n}^{t_{n+1}} \varphi(t, y(t))\,dt \approx h\varphi\left(t_n + \frac{h}{2}, y\left(t_n + \frac{h}{2}\right)\right)$$

et on obtient

$$\begin{cases} u_0 = y(t_0) = y_0, \\ u_{n+1} = u_n + h\varphi\left(t_n + \frac{h}{2}, u_{n+1/2}\right) \quad n = 0, 1, 2, \dots N-1 \end{cases}$$

où $u_{n+1/2}$ est une approximation de $y(t_n + h/2)$. Nous pouvons utiliser une prédiction d'EULER progressive sur un demi-pas 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

$$\begin{cases} u_0 = y(t_0) = y_0, \\ \tilde{u}_{n+1/2} = u_n + (h/2)\varphi(t_n, u_n), \\ u_{n+1} = u_n + h\varphi\left(t_n + \frac{h}{2}, \tilde{u}_{n+1/2}\right) \quad n = 0, 1, 2, \dots N-1 \end{cases} \tag{5.5}$$

Il s'agit d'un schéma à 1 pas explicite car il permet d'expliciter u_{n+1} en fonction de u_n.

```
1;

function [t,u]=eulermodifie(t0,T,y0,N,phi)
  t=linspace(t0,T,N);
  u=zeros(N,1);
  u(1)=y0;
  h=t(2)-t(1);
  for n=1:N-1
     utemp=u(n)+h/2*phi(t(n),u(n));
     u(n+1)=u(n)+h*phi(t(n)+h/2,utemp);
  end
end

% TEST :
phi=@(t,y)[-y];
[t,u]=euleremodifie(0,10,100,11,phi);
plot(t,u,'o-',t,100*exp(-t))
```

④ Si on remplace une fonction 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), on a

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

En utilisant cette formule pour approcher la fonction $t \mapsto \varphi(t, y(t))$ on a

$$\int_{t_n}^{t_{n+1}} \varphi(t, y(t))\,dt \approx \frac{h}{2}\big(\varphi(t_n, y(t_n)) + \varphi(t_{n+1}, y(t_{n+1}))\big)$$

et on obtient le ***schéma du trapèze ou de* CRANK-NICOLSON**

$$\begin{cases} u_0 = y(t_0) = y_0, \\ u_{n+1} - \frac{h}{2}\varphi(t_{n+1}, u_{n+1}) = u_n + \frac{h}{2}\varphi(t_n, u_n) \quad n = 0, 1, 2, \dots N-1 \end{cases} \tag{5.6}$$

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.

```
1;

function [t,u]=cranknicolson(t0,T,y0,N,phi)
  t=linspace(t0,T,N);
  u=zeros(N,1);
  u(1)=y0;
  h=t(2)-t(1);
  for n=1:N-1
     u(n+1)=fzero(@(y) -y+u(n)+(h/2)*(phi(t(n),u(n))+phi(t(n+1),y)), u(n));
  end
end

% TEST :
phi=@(t,y)[-y];
[t,u]=cranknicolson(0,10,100,11,phi);
plot(t,u,'o-',t,100*exp(-t))
```

⑤ Pour éviter le calcul implicite de u_{n+1} dans le schéma du trapèze, nous pouvons utiliser une prédiction 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

$$\begin{cases} u_0 = y(t_0) = y_0, \\ \tilde{u}_{n+1} = u_n + h\varphi(t_n, u_n), \\ u_{n+1} = u_n + \frac{h}{2}\left(\varphi(t_n, u_n) + \varphi(t_{n+1}, \tilde{u}_{n+1})\right) \quad n = 0, 1, 2, \dots N-1 \end{cases} \tag{5.7}$$

Il s'agit à nouveau d'un schéma à 1 pas explicite.

```
1;

function [t,u]=heun(t0,T,y0,N,phi)
  t=linspace(t0,T,N);
  u=zeros(N,1);
  u(1)=y0;
  h=(T-t0)/N;
  for n=1:N-1
     u(n+1) = u(n)+(h/2)*( phi(t(n),u(n)) + phi( t(n+1), u(n)+h*phi(t(n),u(n)) ));
  end
end

% TEST :
phi=@(t,y)[-y];
[t,u]=heun(0,10,100,11,phi);
plot(t,u,'o-',t,100*exp(-t))
```

Nom	**Points d'interpolation**	**Polynôme p**	$\int_{t_n}^{t_{n+1}} \varphi(t,y(t))\,\mathrm{d}t \simeq \int_{t_n}^{t_{n+1}} p(t)\,\mathrm{d}t$	**Schéma**
Euler Explicite	t_n	$p(t)=\varphi(t_n,y_n)$	$h\varphi(t_n,y_n)$	$\begin{cases} u_0=y(t_0)=y_0, \\ u_{n+1}=u_n+h\varphi(t_n,u_n) \end{cases}$
Euler Implicite	t_{n+1}	$p(t)=\varphi(t_{n+1},y_{n+1})$	$h\varphi(t_{n+1},y_{n+1})$	$\begin{cases} u_0=y(t_0)=y_0, \\ u_{n+1}=u_n+h\varphi(t_{n+1},u_{n+1}) \end{cases}$
Euler modifié	$t_n+\frac{h}{2}$	$p(t)=\varphi\left(t_n+\frac{h}{2},y_{n+1/2}\right)$	$h\varphi\left(t_n+\frac{h}{2},y_{n+1/2}\right)$	$\begin{cases} u_0=y(t_0)=y_0, \\ \tilde{u}_{n+1/2}=u_n+\frac{h}{2}\varphi(t_n,u_n), \\ u_{n+1}=u_n+h\varphi\left(t_n+\frac{h}{2},\tilde{u}_{n+1/2}\right) \end{cases}$
Trapèze ou Crank-Nicolson	t_n et t_{n+1}	$p(t)=\frac{\varphi(t_{n+1},y_{n+1})-\varphi(t_n,y_n)}{t_{n+1}-t_n}(t-t_n)+\varphi(t_n,y_n)$	$\frac{h}{2}\left(\varphi(t_n,y_n)+\varphi(t_{n+1},y_{n+1})\right)$	$\begin{cases} u_0=y(t_0)=y_0, \\ u_{n+1}=u_n+\frac{h}{2}\varphi(t_n,u_n)+\frac{h}{2}\varphi(t_{n+1},u_{n+1}) \end{cases}$
Heun	t_n et t_{n+1}	$p(t)=\frac{\varphi(t_{n+1},y_{n+1})-\varphi(t_n,y_n)}{t_{n+1}-t_n}(t-t_n)+\varphi(t_n,y_n)$	$\frac{h}{2}\left(\varphi(t_n,y_n)+\varphi(t_{n+1},y_{n+1})\right)$	$\begin{cases} u_0=y(t_0)=y_0, \\ \tilde{u}_{n+1}=u_n+h\varphi(t_n,u_n), \\ u_{n+1}=u_n+\frac{h}{2}\varphi(t_n,u_n)+\frac{h}{2}\varphi(t_{n+1},\tilde{u}_{n+1}) \end{cases}$

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 (5.1) 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 $|q| < 1$.

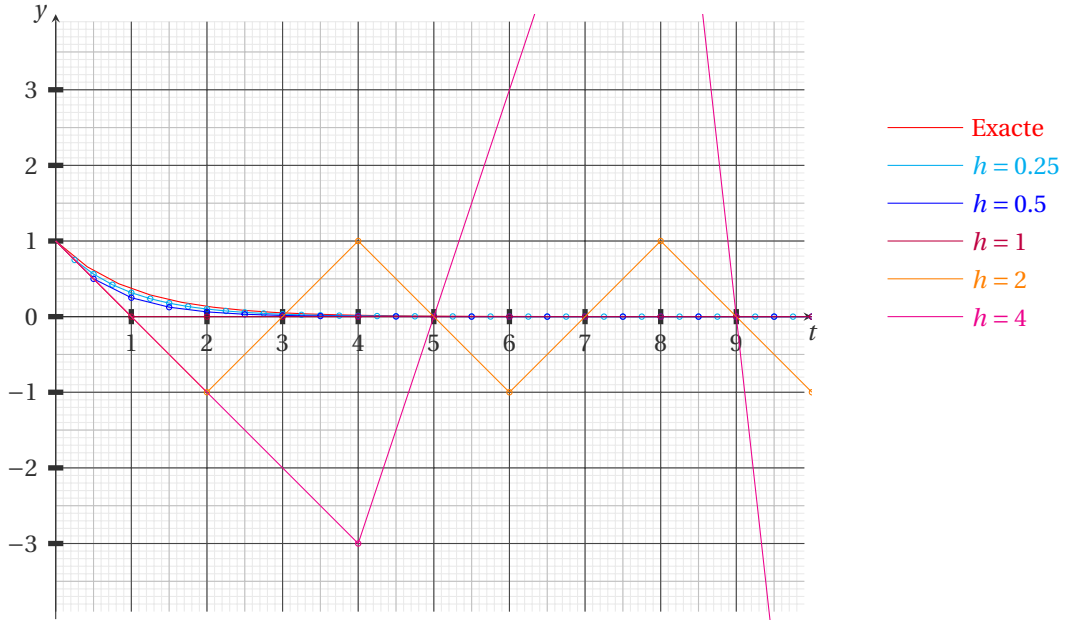
   De la formule $u_{n+1} = (1-h)^{n+1}$ on déduit que
   - ⋆ si $0 < h < 1$ alors la solution numérique est stable et convergente,
   - ⋆ si $h = 1$ alors la solution numérique est stationnaire $u_n = 0$ pour tout $n \in \mathbb{N}^*$,
   - ⋆ si $1 < h < 2$ alors la solution numérique oscille mais est 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

$$u_{n+1}=\frac{1}{(1+h)^{n+1}}.$$

Voyons ce que cela donne avec différentes valeurs de $h>0$:

⋆ si $h=4$ alors $t_n=4n$ et $u_n=\left(\frac{1}{5}\right)^n$ tandis que $y(t_n)=e^{-4n}$,

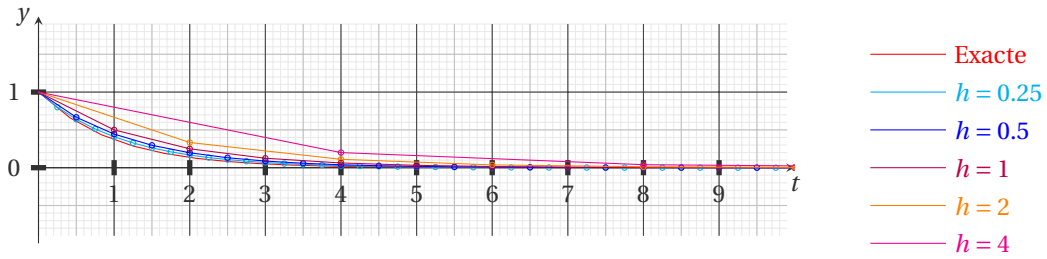
⋆ si $h=2$ alors $t_n=2n$ et $u_n=\left(\frac{1}{3}\right)^n$ tandis que $y(t_n)=e^{-2n}$,

⋆ si $h=1$ alors $t_n=n$ et $u_n=\left(\frac{1}{2}\right)^n$ tandis que $y(t_n)=e^{-n}$,

⋆ si $h=\frac{1}{2}$ alors $t_n=n/2$ et $u_n=\left(\frac{2}{3}\right)^n$ tandis que $y(t_n)=e^{-n/2}$,

⋆ si $h=\frac{1}{4}$ alors $t_n=n/4$ et $u_n=\left(\frac{4}{5}\right)^n$ 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 implicite.



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 $q = 1/(1+h) \in]0;1[$.

5.3.2. Schémas d'Adam

Si nous intégrons l'EDO $y'(t) = \varphi(t, y(t))$ entre t_n et t_{n+1} nous obtenons

$$y_{n+1} - y_n = \int_{t_n}^{t_{n+1}} \varphi(t, y(t)) dt.$$

On peut construire différentes schémas selon la formule de quadrature utilisée pour approcher le membre de droite. Cette solution approchée sera obtenue en construisant une suite récurrente comme suit :

$$\begin{cases} u_0 = y_0, \\ u_{n+1} = u_n + \displaystyle\int_{t_n}^{t_{n+1}} \text{un polynôme d'interpolation de } \varphi(t,u)\ \mathrm{d}t. \end{cases}$$

Les schémas d'Adam approchent l'intégrale $\int_{t_n}^{t_{n+1}} \varphi(t, y(t))\ \mathrm{d}t$ par l'intégrale d'un polynôme p interpolant φ en des points donnés qui peuvent être à l'extérieur de l'intervalle $[t_n; t_{n+1}]$. On peut construire différentes schémas selon les points d'interpolation choisis. Ils se divisent en deux familles : les méthodes d'Adam-Bashforth qui sont explicites et les méthodes d'Adam-Moulton qui sont implicites :

schémas AB-q : les schémas d'Adam-Bashforth d'ordre q approchent l'intégrale $\int_{t_n}^{t_{n+1}} \varphi(t, y(t))\ \mathrm{d}t$ par l'intégrale $\int_{t_n}^{t_{n+1}} p(t)\ \mathrm{d}t$ où p est le polynôme interpolant φ en t_{n-i} pour $0 \le i \le q-1$;

schémas AM-q : les schémas d'Adam-Moulton d'ordre q approchent l'intégrale $\int_{t_n}^{t_{n+1}} \varphi(t, y(t))\ \mathrm{d}t$ par l'intégrale $\int_{t_n}^{t_{n+1}} p(t)\ \mathrm{d}t$ où p est le polynôme interpolant φ en t_{n+1-i} pour $0 \le i \le q$.

Notons qu'il est donc possible de calculer successivement u_q, u_{q+1},..., en partant de $u_0, u_1, \ldots, u_{q-1}$ (qui doivent donc être initialisés par des approximations adéquates car seul u_0 est donné).

Exemple (AB-1)
On a

$$p(t) = \varphi(t_n, y(t_n))$$
$$\int_{t_n}^{t_{n+1}} p(t)\ \mathrm{d}t = h\varphi(t_n, y(t_n))$$

et on obtient le schéma

$$\begin{cases} u_0 = y(t_0) = y_0, \\ u_{n+1} = u_n + h\varphi(t_n, u_n) \quad n = 0, 1, \ldots N-1 \end{cases}$$

La méthode AB_1 coïncide avec la méthode d'Euler progressive.

EXEMPLE (AB-2)
On a

$$p(t) = \frac{\varphi(t_n, y(t_n)) - \varphi(t_{n-1}, y(t_{n-1}))}{h}(t - t_{n-1}) + \varphi(t_{n-1}, y(t_{n-1}))$$
$$\int_{t_n}^{t_{n+1}} p(t)\,\mathrm{d}t = \frac{h}{2}\left(3\varphi(t_n, y(t_n)) - \varphi(t_{n-1}, y(t_{n-1}))\right)$$

et on obtient le schéma

$$\begin{cases} u_0 = y(t_0) = y_0, \\ u_1 = u_0 + h\varphi(t_0, u_0) \approx y(t_1) \\ u_{n+1} = u_n + \frac{h}{2}\left(3\varphi(t_n, u_n) - \varphi(t_{n-1}, u_{n-1})\right) \quad n = 1, 2, \dots N-1 \end{cases}$$

où u_1 est une approximation de $y(t_1)$ obtenue en utilisant une prédiction AB_1.

EXEMPLE (AB-3)
On a

$$p(t) = \frac{\varphi(t_{n-2}, y(t_{n-2}))}{2h^2}(t - t_{n-1})(t - t_n) - \frac{\varphi(t_{n-1}, y(t_{n-1}))}{h^2}(t - t_{n-2})(t - t_n) + \frac{\varphi(t_n, y(t_n))}{2h^2}(t - t_{n-2})(t - t_{n-1})$$
$$\int_{t_n}^{t_{n+1}} p(t)\,\mathrm{d}t = \frac{h}{12}\left(23\varphi(t_n, y(t_n)) - 16\varphi(t_{n-1}, y(t_{n-1})) + 5\varphi(t_{n-2}, y(t_{n-2}))\right)$$

et on obtient le schéma

$$\begin{cases} u_0 = y(t_0) = y_0, \\ u_1 = u_0 + h\varphi(t_0, u_0) \approx y(t_1) \\ u_2 = u_1 + \frac{h}{2}\left(3\varphi(t_1, u_1) - \varphi(t_0, u_0)\right) \approx y(t_2) \\ u_{n+1} = u_n + \frac{h}{12}\left(23\varphi(t_n, u_n) - 16\varphi(t_{n-1}, u_{n-1} + 5\varphi(t_{n-2}, u_{n-2})\right) \quad n = 2, 3, \dots N-1 \end{cases}$$

où u_1 est une approximation de $y(t_1)$ obtenue en utilisant une prédiction AB_1 et u_2 est une approximation de $y(t_2)$ obtenue en utilisant la méthode AB_2.

EXEMPLE (AM-0)
On a

$$p(t) = \varphi(t_{n+1}, y(t_{n+1}))$$
$$\int_{t_n}^{t_{n+1}} p(t)\,\mathrm{d}t = h\varphi(t_{n+1}, y(t_{n+1}))$$

et on obtient le schéma

$$\begin{cases} u_0 = y(t_0) = y_0, \\ u_{n+1} = u_n + h\varphi(t_{n+1}, u_{n+1}) \quad n = 0, 1, \dots N-1 \end{cases}$$

La méthode AM_1 coïncide avec la méthode d'EULER régressive.

EXEMPLE (AM-1)
On a

$$p(t) = \frac{\varphi(t_{n+1}, y(t_{n+1})) - \varphi(t_n, y(t_n))}{h}(t - t_n) + \varphi(t_n, y(t_n))$$
$$\int_{t_n}^{t_{n+1}} p(t)\,\mathrm{d}t = \frac{h}{2}\left(\varphi(t_{n+1}, y(t_{n+1})) + \varphi(t_n, y(t_n))\right)$$

et on obtient le schéma

$$\begin{cases} u_0 = y(t_0) = y_0, \\ u_{n+1} = u_n + \frac{h}{2}\left(\varphi(t_n, u_n) + \varphi(t_{n+1}, u_{n+1})\right) \quad n = 1, 2, \dots N-1 \end{cases}$$

La méthode AM_2 coïncide avec la méthode de CRANK-NICOLSON.

EXEMPLE (AM-2)
On a

$$p(t) = \frac{\varphi(t_{n-1}, y(t_{n-1}))}{2h^2}(t-t_n)(t-t_{n+1}) - \frac{\varphi(t_n, y(t_n))}{h^2}(t-t_{n-1})(t-t_{n+1}) + \frac{\varphi(t_{n+1}, y(t_{n+1}))}{2h^2}(t-t_{n-1})(t-t_n)$$
$$\int_{t_n}^{t_{n+1}} p(t)\,\mathrm{d}t = \frac{h}{12}\left(5\varphi(t_{n+1}, y(t_{n+1})) + 8\varphi(t_n, y(t_n)) - \varphi(t_{n-1}, y(t_{n-1}))\right)$$

et on obtient le schéma

$$\begin{cases} u_0 = y(t_0) = y_0, \\ u_1 = u_0 + \frac{h}{2}\left(\varphi(t_1, u_1) + \varphi(t_0, u_0)\right) \\ u_{n+1} = u_n + \frac{h}{12}\left(5\varphi(t_{n+1}, u_{n+1}) + 8\varphi(t_n, u_n) - \varphi(t_{n-1}, u_{n-1}\right) \quad n = 1, 2, \dots N-1 \end{cases}$$

où u_1 est une approximation de $y(t_1)$ obtenue en utilisant une prédiction AM_1.

5.3.3. Schémas de Nyström et de Milne-Simpson

Les méthodes d'Adam peuvent être facilement généralisées en intégrant l'EDO $y'(t) = \varphi(t, y(t))$ entre t_{n-r} et t_{n+1} avec $r \geq 1$. Par exemple, avec $r = 1$ on obtient les schémas de NYSTRÖM, qui sont explicites, et les schémas de MILNE-SIMPSON, qui sont implicites :

schémas N-q : les schémas de NYSTRÖM d'ordre q approchent l'intégrale $\int_{t_{n-1}}^{t_{n+1}} \varphi(t, y(t))\,\mathrm{d}t$ par l'intégrale $\int_{t_{n-1}}^{t_{n+1}} p(t)\,\mathrm{d}t$ où p est le polynôme interpolant φ en t_{n-i} pour $0 \leq i \leq q-1$;

schémas MS-q : les schémas de MILNE-SIMPSON d'ordre q approchent l'intégrale $\int_{t_{n-1}}^{t_{n+1}} \varphi(t, y(t))\,\mathrm{d}t$ par l'intégrale $\int_{t_{n-1}}^{t_{n+1}} p(t)\,\mathrm{d}t$ où p est le polynôme interpolant φ en t_{n+1-i} pour $0 \leq i \leq q$.

EXEMPLE (N-1)
On a

$$p(t) = \varphi(t_n, y(t_n))$$
$$\int_{t_{n-1}}^{t_{n+1}} p(t)\,\mathrm{d}t = 2h\varphi(t_n, y(t_n))$$

et on obtient le schéma

$$\begin{cases} u_0 = y(t_0) = y_0, \\ u_1 = u_0 + h\varphi(t_0, u_0) \approx y(t_1) \\ u_{n+1} = u_{n-1} + 2h\varphi(t_n, u_n) \quad n = 0, 1, \dots N-1 \end{cases}$$

où u_1 est une approximation de $y(t_1)$ obtenue en utilisant une prédiction d'Euler explicite. La méthode N_1 coïncide avec la méthode du point milieu (appelée aussi Saute-mouton ou *Leapfrog*).

EXEMPLE (MS-0)
On a

$$p(t) = \varphi(t_{n+1}, y(t_{n+1}))$$
$$\int_{t_{n-1}}^{t_{n+1}} p(t)\,\mathrm{d}t = 2h\varphi(t_{n+1}, y(t_{n+1}))$$

et on obtient le schéma

$$\begin{cases} u_0 = y(t_0) = y_0, \\ u_1 = u_0 + h\varphi(t_0, u_0) \approx y(t_1) \\ u_{n+1} = u_{n-1} + 2h\varphi(t_{n+1}, u_{n+1}) \quad n = 0, 1, \dots N-1 \end{cases}$$

où u_1 est une approximation de $y(t_1)$ obtenue en utilisant une prédiction d'Euler explicite.

5.4. Exercices

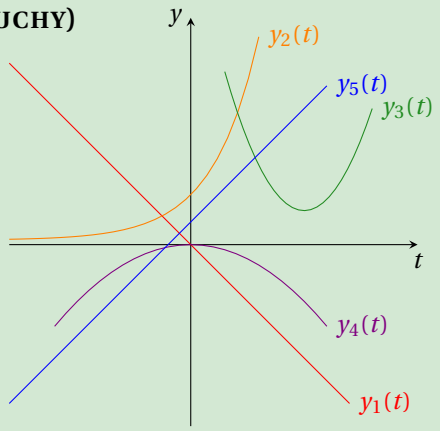
Étude qualitative d'un problème de Cauchy

Exercice 5.1 (Étude qualitative d'un problème de CAUCHY)

On considère l'équation différentielle

$$y'(t) = \frac{e^t}{t^2+1} y(t)$$

Sans résoudre l'équation différentielle, déterminer, parmi les courbes tracées ci-contre, celles qui ne représentent sûrement pas une fonction solution de cette EDO et celles qui sont susceptibles d'en représenter une.



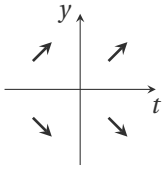
Correction

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 $y(t) = 0$: si la courbe coupe l'axe des abscisses, alors elle a une tangente horizontale en ce point. Les courbes y_2 (orange) et y_3 (verte) ne coupent pas l'axe des abscisses. Les courbes y_1 (rouge), y_4 (violette) et y_5 (bleu) sont les seules courbes qui coupent l'axe des abscisses ; les courbes y_1 (rouge) et y_5 (bleu) n'ayant pas de tangente horizontale en ce point, elles ne conviennent pas.

L'équation impose aussi que $y(t)$ et $y'(t)$ sont de même signe :

Sens de variation



Parmi les courbes restantes, cette condition n'est pas satisfaite par les courbes y_3 (verte) et y_4 (violette).

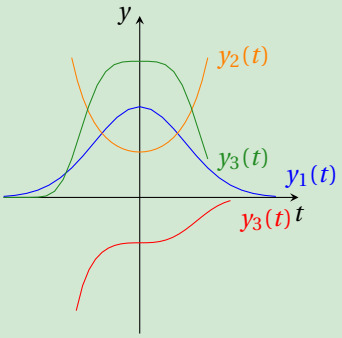
La courbe y_2 (orange) est la seule susceptible de représenter une solution à l'EDO.

Exercice 5.2

Pour $t \in \mathbb{R}$, on considère les quatre équations différentielles

(a) $y'(t) = -ty(t)$
(b) $y'(t) = ty(t)$
(c) $y'(t) = -t^2 y(t)$
(d) $y'(t) = -t^3 y(t)$

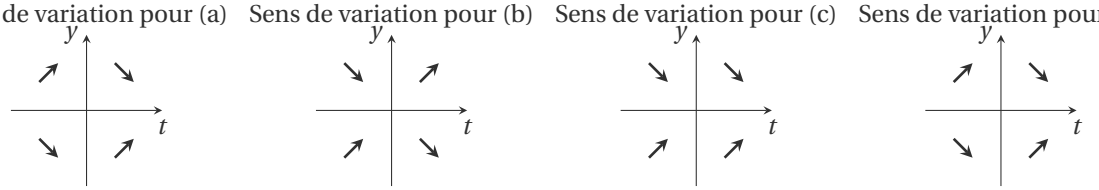
Les graphes de ces fonctions sont tracés sur le graphique à coté. Sans résoudre d'équations différentielles, déterminer pour chaque fonction laquelle des courbes suivantes la représente.



Correction

Pour chaque EDO on décompose le plan cartésien en quatre partie et on trace le sens de variation de sa solution :

Sens de variation pour (a) Sens de variation pour (b) Sens de variation pour (c) Sens de variation pour (d)



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 5.3 (Étude qualitative d'une EDO)

On considère le problème de Cauchy

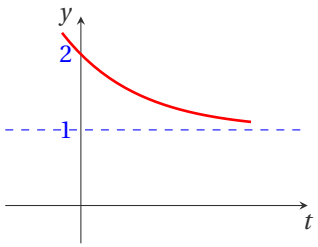
$$\begin{cases} y'(t) = 1 - y(t), & t \in \mathbb{R} \\ y(0) = 2. \end{cases}$$

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 minorée ;
2. étudier la monotonie de la solution ;
3. calculer la limite pour $t \to +\infty$ 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. $y(t) = 1$ 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
$$y(t) > 1 \quad \forall t \in \mathbb{R}.$$
2. 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.
3. La solution est décroissante et minorée donc les limites existent et $\lim_{t\to+\infty} y(t) = \ell \geq 1$. 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 $\lim_{t\to+\infty} y'(t) = \lim_{t\to+\infty} 1 - y(t) = \alpha > 0$, *i.e.* y a une asymptote oblique en $+\infty$, ce qui n'est pas possible.
Par conséquent $\lim_{t\to+\infty} y(t) = 1$.
4. $y''(t) = (y'(t))' = (1 - y(t))' = -y'(t) = y(t) - 1$.
5. Comme $y(t) > 1$ pour tout $t \in \mathbb{R}$, alors $y''(t) > 0$ pour tout $t \in \mathbb{R}$: la solution est convexe.
6. Graphe de la solution :



Exercice 5.4 (Étude qualitative d'une EDO)

On considère le problème de CAUCHY

$$\begin{cases} y'(t) = 4 - y^2(t), \\ y(0) = 0. \end{cases}$$

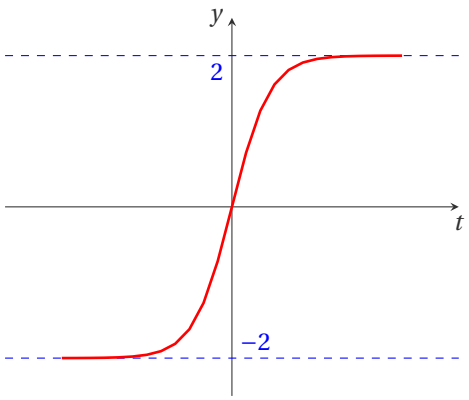
Supposons que le problème admet une et une seule solution $t \mapsto y(t)$ continue et définie sur $\mathbb{R}$.

1. Montrer que la solution est bornée et calculer ces bornes ;
2. étudier la monotonie de la solution ;
3. calculer les limites pour $t \to \pm\infty$ 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. 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
$$y(t) \in]-2;2[\quad \forall t \in \mathbb{R}.$$
2. 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.

3. La solution est croissante et bornée donc les limites existent et $\lim_{t\to+\infty} y(t) = \ell \leq 2$. Si $\ell < 2$ alors $\lim_{t\to+\infty} y'(t) = \lim_{t\to+\infty} 4 - y^2(t) = 4 - \ell^2 > 0$, *i.e.* y a une asymptote oblique en $+\infty$, ce qui n'est pas possible car y est bornée. Par conséquent $\lim_{t\to+\infty} y(t) = 2$.
   Avec le même type de raisonnement on prouve que $\lim_{t\to-\infty} y(t) = -2$.
4. $y''(t) = (y'(t))' = (4 - y^2(t))' = -2y(t)y'(t) = -2y(t)(4 - y^2(t))$.
5. 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 $t > 0$.
6. Graphe de la solution :



Calcul analytique des solutions d'une EDO d'ordre 1 à variables séparables

Exercice 5.5
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 :

$$y'(x) + 2xy^2(x) = 0 \implies \frac{y'(x)}{y^2(x)} = -2x \implies \int y^{-2}\,\mathrm{d}y = -2\int x\,\mathrm{d}x \implies y(x) = \frac{1}{x^2 + c},\ c \in \mathbb{R}.$$

En imposant la CI on obtient $2 = 1/C$ d'où l'unique solution du problème de Cauchy : $y(x) = \frac{2}{2x^2+1}$.

Exercice 5.6
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 :

$$y'(x) - 4xy^2(x) = 0 \implies \frac{y'(x)}{y^2(x)} = 4x \implies \int y^{-2}\,\mathrm{d}y = 4\int x\,\mathrm{d}x \implies y(x) = \frac{1}{-2x^2 + c},\ c \in \mathbb{R}.$$

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 5.7
Résoudre le problème de Cauchy

$$\begin{cases} y'(t) = t y^2(t), \\ y(0) = y_0, \end{cases}$$

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 séparable.
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 $y \in \mathbb{R}$; 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

$$\frac{y'(t)}{y^2(t)} = t \implies \frac{1}{y^2}\,\mathrm{d}y = t\,\mathrm{d}t \implies \int \frac{1}{y^2}\,\mathrm{d}y = \int t\,\mathrm{d}t \implies -\frac{1}{y} = \frac{t^2}{2} + C \implies y(x) = -\frac{1}{\frac{t^2}{2} + C}, \text{ pour } C \in \mathbb{R}.$$

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 $t \in \left]-\sqrt{\frac{2}{y_0}}; \sqrt{\frac{2}{y_0}}\right[$ si $y_0 > 0$ (c'est-à-dire l'intervalle plus large possible qui contient $t = 0$).

Exercice 5.8
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 $\mathscr{C}^1(\mathbb{R})$. 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 donnée.
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

$$\int y^{-2m/(2m+1)}(t)\,\mathrm{d}y = \int 1\,\mathrm{d}t$$

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 $\mathscr{C}^1(\mathbb{R})$ 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 5.9 (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
   $$n(t) = De^{-kt}, \qquad D \in \mathbb{R}^+.$$
   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 $n_0/2 = n(5700) = n_0 e^{-5700t} k$, on obtient $k = \ln 2^{-5700} \approx 1.216\, 10^{-4}$.
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 5.10

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

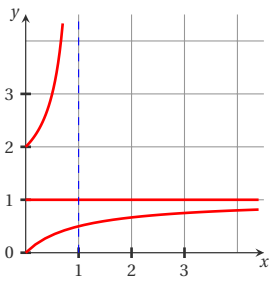
$$y'(t) = (1 - y(t))^2 \qquad \text{pour tout } t \geq 0.$$

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 $\lim_{t\to+\infty} y(t) = \ell$ existe. Puis, en admettant que $\lim_{t\to+\infty} y'(t) = 0$, déterminer ℓ.
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
   $$y'(t) = (1 - y(t))^2 \geq 0 \qquad \text{pour tout } t \geq 0$$
   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 $t \to +\infty$. On note $\lim_{t\to+\infty} y(t) = \ell$. On suppose que $\lim_{t\to+\infty} y'(t) = \ell$. En passant à la limite dans l'EDO on obtient :
   $$0 = (1-\ell)^2$$
   soit $\ell = 1$.
5. ⋆ Si $y(0) = 1$ on sait que $y(t) = 1$ pour tout $t > 0$.
   ⋆ Si $y(0) = 0$ on sait que la fonction y est croissante et $\lim_{t\to+\infty} y'(t) = 1$. En effet, il s'agit d'une EDO à variables séparables et on peut écrire
   $$\int (1-y)^{-2}\,\mathrm{d}y = t, \quad \textit{i.e.} \quad y(t) = \frac{t+c-1}{t+c}$$
   qui existe sur $]-\infty; -c[\cup]-c; +\infty[$, d'où, en imposant $y(0) = 0$, la solution
   $$y(t) = \frac{t}{1+t}, \quad \forall t \geq 0.$$
   ⋆ 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 5.11 (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 augment (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 $p(t) = A$:

$$0 = A(k - hA) \qquad \Longleftrightarrow \qquad A = 0 \text{ ou } A = \frac{k}{h}.$$

On trouve ainsi deux solutions constantes :

$$p(t) \equiv 0 \qquad \text{et} \qquad p(t) \equiv \frac{k}{h}.$$

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

$$\int \frac{\mathrm{d}p}{p(k-hp)} = \int 1 \,\mathrm{d}t$$

i.e.

$$\frac{1}{k}\int \frac{\mathrm{d}p}{p} + \int \frac{h}{k-hp}\,\mathrm{d}p = \int 1 \,\mathrm{d}t.$$

On obtient

$$\frac{1}{k}\ln\frac{|p|}{|k-hp|} = (t+C)$$

et en on déduit

$$p(t) = \frac{kDe^{kt}}{1+hDe^{kt}}.$$

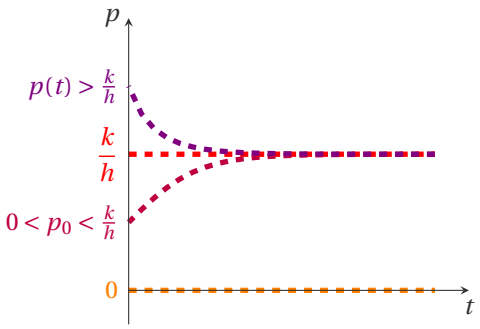
En imposant la condition initiale $p(0) = p_0$ on trouve la constante d'intégration D :

$$D = \frac{p_0}{k-hp_0} = \frac{1}{\frac{k}{p_0}-h}.$$

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 $\lim_{t\to+\infty} p(t) = \frac{k}{h}$: une population qui évolue à partir de p_0 individus à l'instant initiale 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 5.12 («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 de 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 5 000 individus dont 160 sont malades et 1 200 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 positive 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

$$\begin{aligned}
&I'(t) = kI(t)(5000 - I(t)) && \Longrightarrow && \frac{I'(t)}{I(t)(5000-I(t))} = k && \Longrightarrow \\
&\frac{\mathrm{d}I}{I(5000-I)} = k\,\mathrm{d}t && \Longrightarrow && \int \frac{1}{I(5000-I)}\mathrm{d}I = k\int \mathrm{d}t && \Longrightarrow \\
&\int \frac{1}{I}\mathrm{d}I - \int \frac{1}{5000-I}\mathrm{d}I = 5000k\int \mathrm{d}t && \Longrightarrow && \ln(I) + \ln(5000-I) = 5000kt + c && \Longrightarrow \\
&\ln\frac{I}{5000-I} = 5000kt + c && \Longrightarrow && \frac{I}{5000-I} = De^{5000kt} && \Longrightarrow \\
&I(t) = \frac{5000De^{5000kt}}{1+De^{5000kt}} && \Longrightarrow && I(t) = \frac{5000}{De^{-5000kt}+1}
\end{aligned}$$

3. La valeur numérique de la constante d'intégration D est obtenue grâce à la CI :

$$160 = I(0) = \frac{5000}{De^0+1} \quad \Longrightarrow \quad 160 = \frac{5000}{1+D} \quad \Longrightarrow \quad D = \frac{4}{121} \quad \Longrightarrow \quad I(t) = \frac{20000}{4+121e^{-5000kt}}.$$

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 :

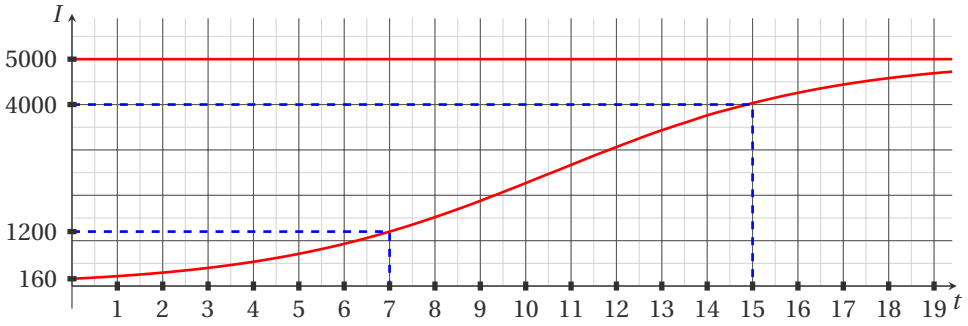
$$1200 = I(7) = \frac{20000}{4+121e^{-35000k}} \quad \Longrightarrow \quad k = \frac{1}{35000}\ln\frac{363}{38} \quad \Longrightarrow \quad I(t) = \frac{20000}{4+121e^{-\frac{t}{7}\ln\left(\frac{363}{38}\right)}}.$$

5. On cherche $\bar{t}$ tel que $I(\bar{t}) = 80\%I_T = \frac{80\times5000}{100} = 4000$:

$$4000 = \frac{20000}{4+121e^{-\frac{\bar{t}}{7}\ln\left(\frac{363}{38}\right)}}$$

d'où $\bar{t} = \frac{1}{5000}\ln(121) \approx 15$ jours.

6. Avec ce modèle $\lim\limits_{t\to+\infty} I(t) = 5000$ mais I ne peut jamais atteindre exactement 100% de la population en un temps fini (deux solution ne s'intersectent jamais).



Exercice 5.13

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 $t \in \mathbb{R}$:

$$0 = 0,022A(20 - A) \qquad \Longleftrightarrow \qquad A = 0 \text{ ou } A = 20.$$

On trouve ainsi deux solutions constantes :

$$y(t) \equiv 0 \qquad \text{et} \qquad y(t) \equiv 20.$$

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

$$\int \frac{\mathrm{d}y}{y(20-y)} = \int 0,022 \, \mathrm{d}t,$$

i.e.

$$\frac{1}{20}\left(\int \frac{\mathrm{d}y}{y} - \int \frac{1}{y-20} \, \mathrm{d}y\right) = \int 0,022 \, \mathrm{d}t.$$

On obtient

$$\ln \frac{|y|}{|y-20|} = 0,44\,t + C \qquad \text{pour tout } C \in \mathbb{R}$$

et on en déduit

$$y(t) = \frac{20}{1 + 20De^{-0,44t}} \qquad \text{pour tout } D \in \mathbb{R}_+.$$

2. Si $t = 0$ correspond à l'année 1980 et si $y(0) = 0,01$ alors

$$0,01 = \frac{20}{1 + 20De^{-0,44\times 0}} \qquad \Longrightarrow \qquad D = 1999$$

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 5.14 (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

1. 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 $t \in \mathbb{R}$:

$$0 = A\ln(A) \qquad \Longleftrightarrow \qquad A = 1.$$

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

$$\int \frac{\mathrm{d}y}{y\ln(y)} = \int -1 \,\mathrm{d}t.$$

On obtient[3]

$$\ln|\ln(y(t))| = -t + C \qquad \text{pour tout } C \in \mathbb{R}$$

et on en déduit

$$y(t) = e^{De^{-t}} \qquad \text{pour tout } D \in \mathbb{R}.$$

2. Si $y(0) > 1$ alors $y'(t) < 0$ (la population décroît) ; si $0 < y(0) < 1$ alors $y'(t) > 0$ (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 5.15
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

$$\big[(x+1)y(x)\big]' = (x+1)\sin(x).$$

En intégrant par parties, on obtient (attention, la constante dépend de l'intervalle)

$$(x+1)y(x) = \int (x+1)\sin(x)\,\mathrm{d}x = -(x+1)\cos(x) + \sin(x) + C.$$

La solution générale sur I_1, ou sur I_2, est donc

$$y(x) = -\cos(x) + \frac{\sin(x) + C}{(x+1)} \qquad \text{avec } C \in \mathbb{R}.$$

Exercice 5.16
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

- $A(x) = \int \frac{3x^2+1}{1}\,\mathrm{d}x = x^3 + x$,
- $B(x) = \int \frac{x^2 e^{-x}}{1} e^{A(x)}\,\mathrm{d}x = \int x^2 e^{x^3}\,\mathrm{d}x = \frac{1}{3}\int 3x^2 e^{x^3}\,\mathrm{d}x = \frac{1}{3}\int e^{u(x)}u'(x)\,\mathrm{d}x = \frac{1}{3}e^{u(x)} = \frac{e^{x^3}}{3}$.

Toutes les solutions de l'EDO sont donc les fonctions $y(x) = \left(C + \frac{e^{x^3}}{3}\right)e^{-x^3-x}$ 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 $y(x) = \left(\frac{2}{3} + \frac{e^{x^3}}{3}\right)e^{-x^3-x}$.

3. $\int \frac{1}{x\ln(x)}\,\mathrm{d}x = \int \frac{1}{z}\,\mathrm{d}z = \ln|z| + c = \ln|\ln(x)| + C$

Exercice 5.17
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

- $A(x) = \int \frac{3x^2-1}{1}\, \mathrm{d}x = x^3 - x$,
- $B(x) = \int \frac{x^2 e^x}{1} e^{A(x)}\, \mathrm{d}x = \int x^2 e^{x^3}\, \mathrm{d}x = \frac{e^{x^3}}{3}$.

Toutes les solutions de l'EDO sont donc les fonctions $y(x) = \left(C + \frac{e^{x^3}}{3}\right) e^{-x^3+x}$ 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 $y(x) = \left(-\frac{4}{3} + \frac{e^{x^3}}{3}\right) e^{-x^3+x}$.

Exercice 5.18
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

- $A(x) = \int \frac{1}{x-1}\, \mathrm{d}x = \ln(1-x)$,
- $B(x) = \int \frac{(x-2)^2}{x-1} e^{A(x)}\, \mathrm{d}x = -\int (x-2)^2\, \mathrm{d}x = \frac{(x-2)^3}{3}$.

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 5.19
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

- $A(x) = \int 4x^3 + 5\, \mathrm{d}x = x^4 + 5x$,
- $B(x) = \int x^3 e^{-5x} e^{A(x)}\, \mathrm{d}x = -\int x^3 e^{x^4}\, \mathrm{d}x = \frac{e^{x^4}}{4}$.

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 5.20
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

- $A(x)=\int 2\,\mathrm{d}x=2x$,
- $B(x)=\int e^{-2x}e^{A(x)}\,\mathrm{d}x=\int 1\,\mathrm{d}x=x$.

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 5.21
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

- $A(x)=\int\frac{b(x)}{a(x)}\,\mathrm{d}x=\int 2x\,\mathrm{d}x=x^2$,
- $B(x)=\int\frac{g(x)}{a(x)}e^{A(x)}\,\mathrm{d}x=\int xe^{A(x)}\,\mathrm{d}x=\int xe^{x^2}\,\mathrm{d}x=\frac{1}{2}e^{x^2}$,

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 5.22
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

- $A(t)=\int\frac{R}{L}\,\mathrm{d}t=\frac{R}{L}t$,
- $B(t)=\int\frac{V}{L}e^{A(t)}\,\mathrm{d}t=\frac{V}{L}\int e^{\frac{R}{L}t}\,\mathrm{d}t=\frac{V}{L}\frac{L}{R}\int\frac{R}{L}e^{-\frac{R}{L}t}\,\mathrm{d}t=\frac{V}{R}e^{\frac{R}{L}t}$,

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 $I(t)\xrightarrow[t\to+\infty]{}\frac{V}{R}$.

Exercice 5.23 («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

- ⋆ $A(t) = \int -\gamma \, \mathrm{d}t = -\gamma t$,
- ⋆ $B(t) = \int -\gamma T_{\text{ext}} e^{A(t)} \, \mathrm{d}t = T_{\text{ext}} \int -\gamma e^{-\gamma t} \, \mathrm{d}t = T_{\text{ext}} e^{-\gamma t}$,

donc toutes les solutions de l'EDO sont les fonctions $T(t) = De^{\gamma t} + T_{\text{ext}}$ pour $D \in \mathbb{R}$.
La valeur numérique de la constante d'intégration D est obtenue grâce à la CI : $T_0 = T(0) = T_{\text{ext}} + De^{\gamma \cdot 0}$ donc $D = T_0 - T_{\text{ext}}$.
Ici $T_{\text{ext}} = -5°\text{C}$ et $T_0 = 20°\text{C}$ donc la température du cadavre suit la loi

$$T(t) = -5 + 25e^{\gamma t}.$$

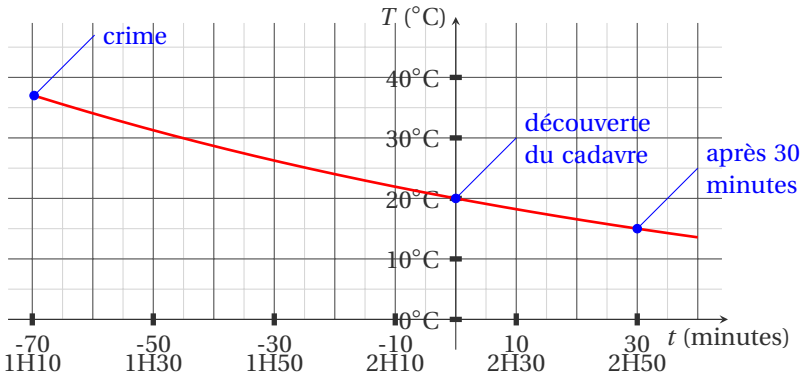
On sait que $15 = T(30) = -5 + 25e^{30\gamma}$ d'où $\gamma = \frac{\ln(4/5)}{30}$. La température du corps suit donc la loi

$$T(t) = -5 + 25e^{\frac{\ln(4/5)}{30} t}.$$

Pour déterminer l'heure du meurtre il faut donc résoudre l'équation

$$37 = -5 + 25e^{\frac{\ln(4/5)}{30} t}$$

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 5.24 («Un gâteau presque parfait»)
Un gâteau est sorti du four à 17H00 quand il est brûlant (100°C). Après 10 minutes sa température est de 80°C et de 65°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

- ⋆ $A(t) = \int -\gamma \, \mathrm{d}t = -\gamma t$,
- ⋆ $B(t) = \int -\gamma T_{\text{ext}} e^{A(t)} \, \mathrm{d}t = T_{\text{ext}} \int -\gamma e^{-\gamma t} \, \mathrm{d}t = T_{\text{ext}} e^{-\gamma t}$,

donc toutes les solutions de l'EDO sont les fonctions $T(t) = De^{\gamma t} + T_{\text{ext}}$ pour $D \in \mathbb{R}$.
La valeur numérique de la constante d'intégration D est obtenue grâce à la CI :

$$T_0 = T(0) = T_{\text{ext}} + De^{\gamma \cdot 0} \quad \Longrightarrow \quad D = T_0 - T_{\text{ext}} \quad \Longrightarrow \quad T(t) = T_{\text{ext}} + (T_0 - T_{\text{ext}})e^{\gamma t}.$$

Ici l'inconnue est T_{ext}. On sait que $T(t = 0) = 100°\text{C}$ et $T(t = 10) = 80°\text{C}$ et $T(t = 20) = 65°\text{C}$. Il s'agit donc de résoudre le système de trois équations en les trois inconnues $\gamma, D, T_{\text{ext}}$:

$$\begin{cases} 100 &= T_{\text{ext}} + D, \\ 80 &= T_{\text{ext}} + De^{10\gamma}, \\ 65 &= T_{\text{ext}} + De^{20\gamma}. \end{cases}$$

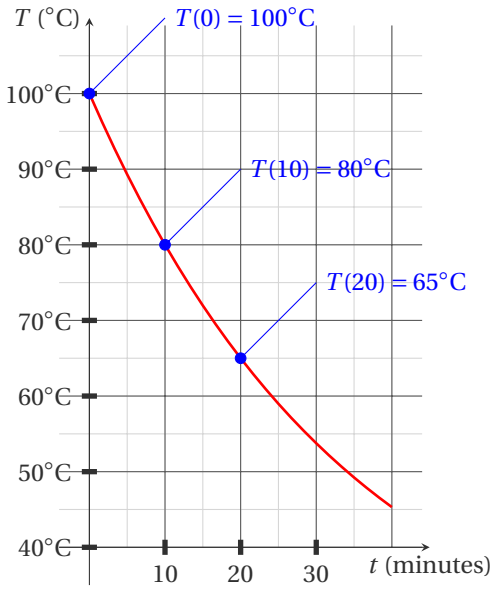
La première équation se réécrit $D = 100 - T_{\text{ext}}$, la seconde

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

$$\frac{80-T_{\text{ext}}}{100-T_{\text{ext}}} = e^{10K} = \frac{e^{20K}}{e^{10K}} = \frac{65-T_{\text{ext}}}{80-T_{\text{ext}}}$$

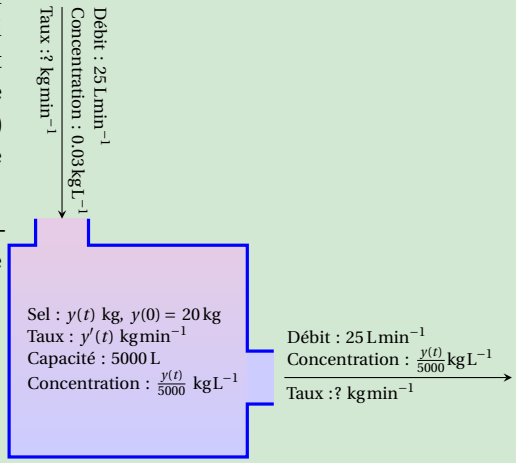
d'où $(80-T_{\text{ext}})^2 = (65-T_{\text{ext}})(100-T_{\text{ext}})$, ainsi $T_{\text{ext}} = \frac{65\times100-80^2}{5} = 20$. La cuisine est donc à 20°C et, plus généralement, la température du gâteau évolue selon la loi

$$T(t) = 20 + 80e^{\frac{\ln(3/4)}{10}t}.$$



Exercice 5.25

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 25 L min^{-1}. La solution est maintenue bien mélangés. 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.



1. Après avoir calculé les taux auxquels le sel entre et sort du réservoir, montrer que cette situation est décrite par le problème de Cauchy
$$\begin{cases} y'(t) = 0.75 - \frac{y(t)}{200}, \\ y(0) = 20. \end{cases}$$
2. Calculer l'unique solutions de ce problème.
3. Combien de sel reste dans le réservoir après une demi-heure?

Correction

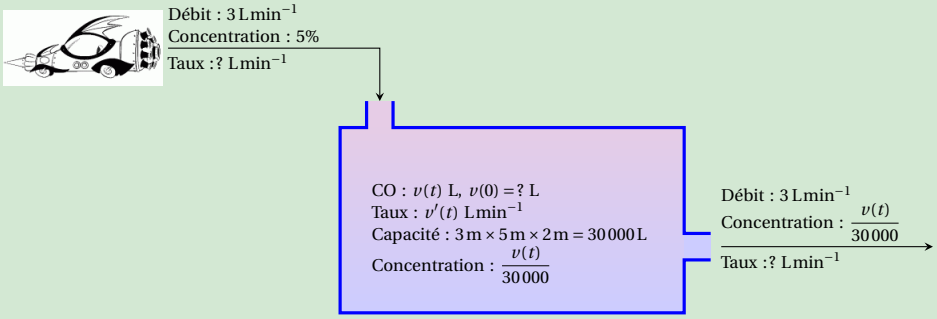
1. Le taux auquel le sel entre est $(0.03\,\text{kg})(25\,\text{L min}^{-1})$=$0.75\,\text{kg min}^{-1}$. Comme le réservoir contient constamment 5000 L de liquide, la concentration est égale à $y(t)/5000$ (exprimée en kg L^{-1}). Le débit du mélange qui sort est alors de $25\,\text{L min}^{-1}$, donc le taux auquel le sel sort est $(\frac{y(t)}{5000}\text{kg L}^{-1})(25\,\text{L min}^{-1})$=$\frac{y(t)}{200}\text{kg min}^{-1}$. L'équation différentielle qui décrit cette variation s'écrit alors
$$y'(t) = 0.75 - \frac{y(t)}{200}$$
2. On a une EDO linéaire d'ordre 1 avec $a(t) = 1$, $b(t) = 1/200$, $g(t) = 0.75$. On pose
   - ⋆ $A(t) = \int \frac{b(t)}{a(t)}\,\mathrm{d}t = \int \frac{1}{200}\,\mathrm{d}t = \frac{1}{200}t$,
   - ⋆ $B(t) = \int \frac{g(t)}{a(t)}e^{A(t)}\,\mathrm{d}t = 0.75\int e^{t/200}\,\mathrm{d}t = 150e^{t/200}$,

   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
$$y(t) = 150 - 130e^{-t/200}.$$
3. Reste à calculer la quantité de sel après 30 minutes : $y(30) = 150 - 130e^{-3/20} \simeq 38.1\,\text{kg}$.

Exercice 5.26

L'air d'un garage de $3\,\mathrm{m} \times 5\,\mathrm{m} \times 2\,\mathrm{m}$ 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 $3\,\mathrm{L\,min^{-1}}$. On note $v(t)$ le volume de CO présent dans la pièce au temps t.

1. En supposant que le mélange se fait instantanément, montrer que cette situation est décrite par le problème de Cauchy
$$\begin{cases} v'(t) = 0.15 - \frac{v(t)}{10000}, \\ v(0) = 0.3. \end{cases}$$
2. 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.
3. Le seuil critique pour la santé est de 0.015% de CO. Après combien de temps ce taux est-il atteint ?



Correction

1. Le taux de CO produit par minute est $0.05 \times 3\,\mathrm{L\,min^{-1}} = 0.15\,\mathrm{L\,min^{-1}}$. Le débit de l'air qui sort est de $3\,\mathrm{L\,min^{-1}}$, donc le taux auquel le CO sort est $\frac{v(t)}{30000} \times 3\,\mathrm{L\,min^{-1}} = \frac{v(t)}{10000}\mathrm{L\,min^{-1}}$. 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 $0.001\% \times 30000\,\mathrm{L} = 0.3\,\mathrm{L}$.
2. On a une EDO linéaire d'ordre 1 avec $a(t) = 1$, $b(t) = 1/10000$, $g(t) = 0.15$. On pose
   - $A(t) = \int \frac{b(t)}{a(t)}\,\mathrm{d}t = \int \frac{1}{10000}\,\mathrm{d}t = \frac{1}{10000}t$,
   - $B(t) = \int \frac{g(t)}{a(t)} e^{A(t)}\,\mathrm{d}t = 0.15 \int e^{t/10000}\,\mathrm{d}t = 1500 e^{t/10000}$,

   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
$$v(t) = 1500 - (1500 - 0.3)e^{-t/10000} \xrightarrow[t\to+\infty]{} 1500.$$
3. Reste à calculer après combien de minutes le taux de CO atteint 0.015% : $0.00015 = 1500 - (1500 - 0.3)e^{-t/10000}$ ssi $t = 10000 \ln\left(\frac{1500-0.3}{1500-4.5}\right) \simeq 28.04\,\mathrm{min}$.

Exercice 5.27 (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-elle 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 ℓ' 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

$$y(t) = \int_0^t \frac{1}{L(\tau)}\mathrm{d}\tau = \frac{1}{1000}\int_0^t \frac{1}{1+\tau}\mathrm{d}\tau = \frac{1}{1000}\left[\ln(1+\tau)\right]_0^t = \frac{1}{1000}\ln(1+t)$$

et donc

$$\ell(t) = y(t)L(t) = \frac{1000\,t + 1000}{1000}\ln(1+t) = (1+t)\ln(1+t) \quad \forall\, t \geq 0.$$

L'escargot touchera l'extrémité mobile de l'élastique lorsque $\ell(t) = L(t)$, c'est-à-dire à l'instant $t_f = e^{1000} - 1 \simeq 1.97 \times 10^{434}$ jours $\simeq 5.397 \times 10^{431}$ années (ce qui correspond à $\simeq 3.9 \times 10^{421}$ fois l'age de l'univers).
En étudiant la fonction $t \mapsto d(t) = L(t) - \ell(t)$, on trouve que cette distance est maximale [4] 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 5.28
Déterminer la solution générale des EDO suivantes après avoir indiqué sur quelle intervalle la solution est définie :

1. $y'(t) - \frac{1}{t}y(t) = (y(t))^3\sin(t)$
2. $y'(t) + ty(t) = t^3(y(t))^2$

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 $]0;+\infty[$.

- $A(t) = (1-\alpha)\displaystyle\int \frac{v(t)}{u(t)}\,\mathrm{d}t = \int \frac{1}{t}\,\mathrm{d}t = 2\ln|t|$,
- $B(t) = (1-\alpha)\displaystyle\int \frac{w(t)}{u(t)}e^{A(t)}\,\mathrm{d}t = -\int t^2\sin(t)\,\mathrm{d}t = 2t^2\cos(t) - 4t\sin(t) - 4\cos(t)$,
- $z(t) = \big(C_{1,2} + B(t)\big)\,e^{-A(t)} = \big(C_{1,2} + 2t^2\cos(t) - 4t\sin(t) - 4\cos(t)\big)\,e^{-2\ln|t|} = \frac{C_{1,2} + 2t^2\cos(t) - 4t\sin(t) - 4\cos(t)}{t^2}$,
- $y(t) = (z(t))^{-1/2} = \frac{1}{\sqrt{z(t)}}$

et on conclut que la solution générale de l'EDO de Bernoulli assignée est

$$y\colon \mathbb{R}^* \to \mathbb{R}$$
$$t \mapsto \begin{cases} \dfrac{t}{\sqrt{C_1 + 2t^2\cos(t) - 4t\sin(t) - 4\cos(t)}} & \text{si } t < 0 \qquad \text{avec } C_1 \in \mathbb{R}^+, \\ \dfrac{-t}{\sqrt{C_2 + 2t^2\cos(t) - 4t\sin(t) - 4\cos(t)}} & \text{si } t > 0 \qquad \text{avec } C_2 \in \mathbb{R}^+, \end{cases}$$

(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 $\mathbb{R}$.

- *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. $d'(t) = 999 - \ln(1+t)$

⋆ $A(t) = (1-\alpha)\displaystyle\int \frac{v(t)}{u(t)}\,\mathrm{d}t = -\int t\,\mathrm{d}t = -\frac{t^2}{2}$,

⋆ $B(t) = (1-\alpha)\displaystyle\int \frac{w(t)}{u(t)} e^{A(t)}\,\mathrm{d}t = -\int t^3 e^{-t^2/2}\,\mathrm{d}t = -2\int xe^x\,\mathrm{d}x = -2(x-1)e^x = (2+t^2)e^{-t^2/2}$,

⋆ $z(t) = ce^{t^2/2} + 2 + t^2$,

⋆ $y(t) = (z(t))^{-1} = \frac{1}{z(t)}$

et on conclut que la solution générale de l'EDO de BERNOULLI assignée est

$$\begin{aligned} y\colon \mathbb{R} &\to \mathbb{R} \\ t &\mapsto \frac{1}{Ce^{t^2/2} + t^2 + 2} \qquad \text{avec } C \in \mathbb{R} \end{aligned}$$

Approximation numérique d'EDO

☆ Exercice 5.29

Considérons le problème de CAUCHY

trouver une fonction $y\colon I \subset \mathbb{R} \to \mathbb{R}$ 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 $n = 0, 1, 2, \ldots, N$ 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.*
Pour chaque nœud t_n, on cherche la valeur inconnue u_n qui approche la valeur exacte $y(t_n)$. L'ensemble des valeurs $\{u_0 = y_0, u_1, \ldots, u_N\}$ représente la solution numérique.
Tracer, avec les classiques, la solution de l'équation différentielle $y'(t) = \frac{-y(t)}{2(t+1)}$ et $y(0) = 1$ sur l'intervalle $[0;8]$.

Correction
La solution exacte est $y(t) = 1/\sqrt{t+1}$ qu'on va comparer aux solutions obtenues avec les méthodes indiquées dans le cours :

```
t0=0;
T=8;
y0=1;
phi=@(t,y)[-y/(2*(t+1))];

N=10;
[t,uE]=eulerexplicite(t0,T,y0,N,phi);
[t,uI]=eulerimplicite(t0,T,y0,N,phi);
[t,uM]=eulermodifie(t0,T,y0,N,phi);
[t,uCN]=cranknicolson(t0,T,y0,N,phi);
[t,uH]=heun(t0,T,y0,N,phi);
y=1./sqrt(t+1);

plot(t,y,'-',t,uE,'*-',t,uM,'o-',t,uI,'+-',t,uCN,'.-',t,uH,'x-')
legend(['Exacte';'Euler Explicite'; 'Euler Modifiee'; 'Euler Implicite';'Crank-Nicolson';'Heun'])
```

☆ Exercice 5.30

Considérons le problème de CAUCHY

trouver la fonction $y\colon I \subset \mathbb{R} \to \mathbb{R}$ 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, …, 2^{12} (ce qui correspond à différentes valeurs de h, à savoir 1/2, 1/4, 1/8, …, 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 vecteurs `err` de sort

que `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

$$\{(\ln(h_k),\ln(\texttt{err}_k))\}_{k=1}^{N}.$$

En effet, si l'erreur `err` est égale à Ch^p alors $\ln(\texttt{err})=\ln(C)+p\ln(h)$: en échelle logarithmique, p représente donc la pente de la ligne droite $\ln(\texttt{err})$.

Correction
On initialise les données :

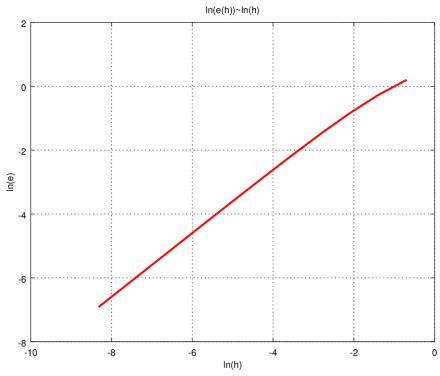
```
t0=0;
T=1;
y0=1;
phi=@(t,y)[y];
exacte=@(t)[exp(t)];
```

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
   [t,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 5.31
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 discrétisation 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

$$u_{n+1} = \frac{u_n}{1+\frac{h}{1+t_{n+1}^2}}.$$

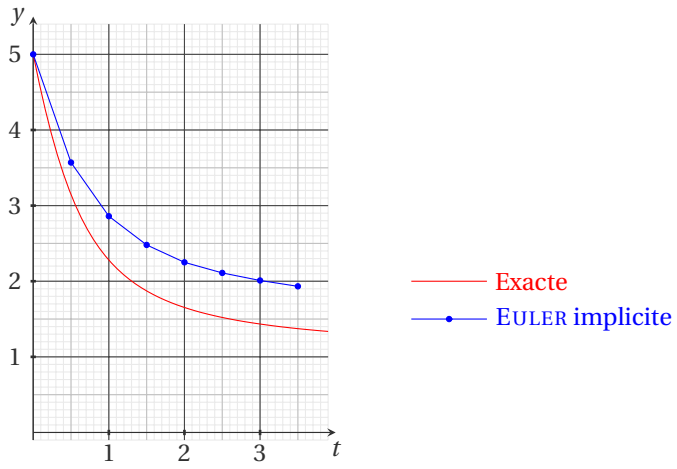
Si à l'instant $t = 0$ la concentration est $y(0) = 5$, et si $h = 1/2$, alors $t_n = n/2$ et

$$u_{n+1} = \frac{4+(n+1)^2}{6+(n+1)^2} u_n.$$

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 5.32 (Loi de NEWTON)
Considérons une tasse de café à la température de 75°C dans une salle à 25°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.

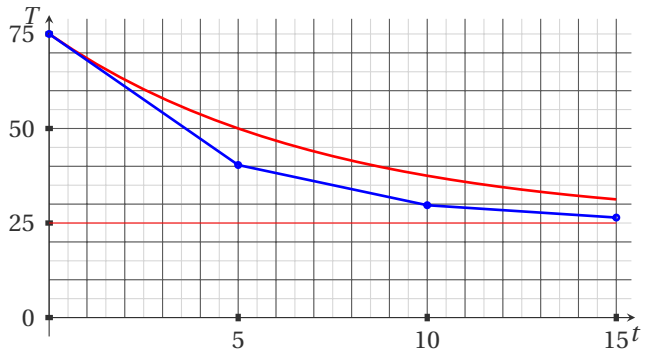
Correction
Supposons de connaître K mais de ne pas vouloir/pouvoir calculer la fonction $T(t)$. Grâce à la méthode d'EULER on peut estimer la température à différentes instantes t_i en faisant une discrétisation temporelle du futur (i.e. on construit une suite de valeurs $\{t_i = 0 + i\Delta t\}_i$) 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

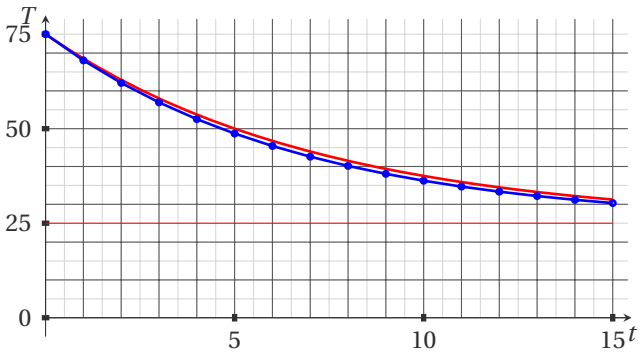
$$\begin{cases} T_{i+1} = (1 - \frac{\ln(2)}{5}\Delta t)\,T_i + 5\ln(2)\Delta t, \\ T_0 = 75. \end{cases}$$

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 5.33 («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 Δt le pas temporel. Écrire le schéma d'EULER implicite pour approcher la solution de cette équation différentielle.
2. Soit $T_{\text{ext}} = 0°\text{C}$. 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 Δt, 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°C,
   - ⋆ à l'heure de la découverte la température du corps est de 20°C,
   - ⋆ la température externe est $T_{\text{ext}} = 0°\text{C}$.

   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 discrétisation Δt, 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 $T_{\text{ext}} = 0°\text{C}$, 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 Δt 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 $t_0 = t_n - n\Delta t = 2H20 - 1H20 = 01H00$.

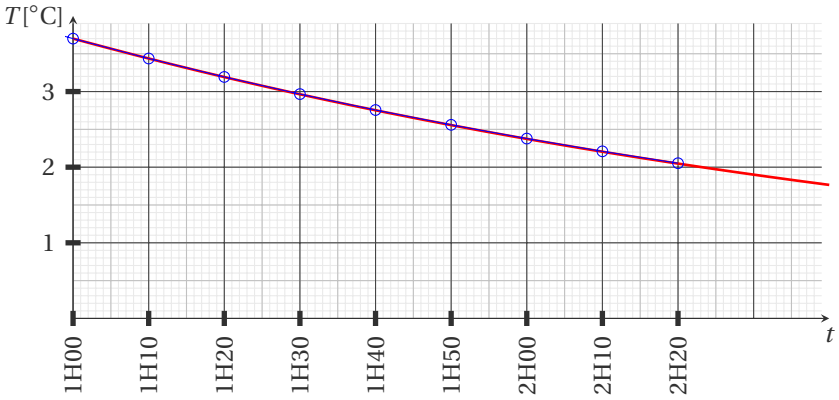
Pour cette équation différentielle, il est possible de calculer analytiquement ses solutions : la température du cadavre suit la loi

$$T(t) = 37e^{Kt}.$$

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 commit à 00H57.



Exercice 5.34
Montrer que le problème de CAUCHY

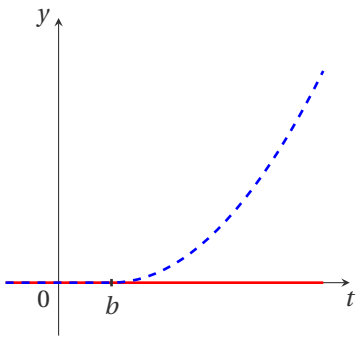
$$\begin{cases} y'(t) = y^{1/2}(t), & t > 0 \\ y(0) = 0, \end{cases}$$

admet une infinité de solutions de classe $\mathscr{C}^1(\mathbb{R}^+)$. Parmi ces solutions, quelle solution approche-t-on si on utilise la méthode d'Euler explicite? et la méthode d'Euler implicite? Que se passe-t-il si la donnée initiale est $y(0) = y_0 > 0$?

Correction
La fonction $\varphi(t,y) = \sqrt{y}$ n'est pas lipschitzienne par rapport à y, donc le théorème d'existence et unicité locale n'est pas valable au voisinage de $(0,0)$. L'EDO est a variables séparables, on peut donc expliciter toutes les solutions du problème de CAUCHY. Elle admet une solution constante, la fonction $y(t) = 0$ pour tout $t \in \mathbb{R}^+$, et des solutions de la forme $y(t) = \frac{1}{4}(t+c)^2$ pour tout $t \geq c$. En imposant la CI on trouve que, pour tout $b \in \mathbb{R}^+$, les fonctions

$$y_b(t) = \begin{cases} 0, & \text{si } 0 \leq t \leq b, \\ \frac{1}{4}(t-b)^2, & \text{si } t \geq b, \end{cases}$$

sont de classe $\mathscr{C}^1(\mathbb{R}^+)$ et sont solution du problème de CAUCHY donné.



La méthode d'Euler explicite construit la suite

$$\begin{cases} u_0 = y_0 = 0, \\ u_{n+1} = u_n + h\varphi(t_n, u_n) = u_n + h u_n^{1/2}, & n = 0,1,2,\dots N-1 \end{cases}$$

par conséquent $u_n = 0$ pour tout n. La méthode d'Euler explicite approche la solution constante $y(t) =$ pour tout $t \in \mathbb{R}^+$.
La méthode d'Euler implicite construit la suite

$$\begin{cases} u_0 = y_0 = 0, \\ u_{n+1} = u_n + h\varphi(t_{n+1}, u_{n+1}) = u_n + h u_{n+1}^{1/2}, & n = 0,1,2,\dots N-1. \end{cases}$$

par conséquent $u_0 = 0$ mais u_1 dépend de la méthode de résolution de l'équation implicite $x = 0 + h\sqrt{x}$. Bien sur $x = 0$ est une solution mais $x = h^2$ est aussi solution. Si le schéma choisi $u_1 = h^2$, alors $u_n > 0$ pour tout $n \in \mathbb{N}^*$.
Notons que le problème de Cauchy avec une CI $y(0) = y_0 > 0$ admet une et une seule solution, la fonction $y(t) = \frac{1}{4}(t - 2\sqrt{y_0})^2$. Dans ce cas, les deux schémas approchent forcement la même solution.

☆ Exercice 5.35
Considérons une population de bactéries. Soit $p(t)$ le nombre d'individus (≥ 0) à l'instant $t \geq 0$. Un modèle qui décrit l'évolution de cette population est l'«équation de la logistique» : soit k et h deux constantes positives, alors $p(t)$ vérifie l'équation différentielle ordinaire (EDO) du premier ordre

$$p'(t) = kp(t) - hp^2(t).$$

On veut calculer $p(t)$ à partir d'un nombre initiale d'individus donné

$$p(0) = p_0 \geq 0.$$

Correction
Solution exacte

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. Il s'agit d'une EDO à variables séparables.

   On cherche d'abord les solutions constantes, c'est-à-dire les solutions du type $p(t) \equiv c$ pour tout $t \in \mathbb{R}^+$:

   $$0 = kc - hc^2.$$

   On a donc deux solutions constantes :

   $$p(t) \equiv 0 \qquad \text{et} \qquad p(t) \equiv \frac{k}{h}.$$

   Étant donné que deux solutions d'une EDO ne s'intersectent jamais, dorénavant on supposera $p(t) \neq 0$ et $p(t) \neq \frac{k}{h}$ pour tout $t \in \mathbb{R}^+$, ainsi

   $$\frac{p'(t)}{kp(t) - hp^2(t)} = 1.$$

   Formellement on a

   $$\begin{aligned}
   \frac{\mathrm{d}p}{kp - hp^2} = 1\,\mathrm{d}t \qquad &\Longrightarrow \qquad \int \frac{1}{p(k-hp)}\,\mathrm{d}p = \int 1\,\mathrm{d}t \qquad \Longrightarrow \\
   \frac{1}{k}\int \frac{1}{p}\,\mathrm{d}p - \frac{1}{k}\int \frac{-h}{k-hp}\,\mathrm{d}p = \int 1\,\mathrm{d}t \qquad &\Longrightarrow \qquad \frac{1}{k}\ln(p) - \frac{1}{k}\ln(k-hp) = t + c \qquad \Longrightarrow \\
   \ln\left(\frac{p}{k-hp}\right) = kt + kc \qquad &\Longrightarrow \qquad \frac{p}{k-hp} = De^{kt} \qquad \Longrightarrow \\
   p(t) = \frac{k}{\frac{1}{De^{kt}} + h}&.
   \end{aligned}$$

2. La valeur numérique de la constante d'intégration D est obtenue grâce à la CI :

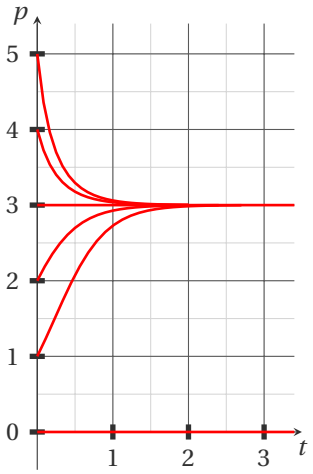
   $$p_0 = p(0) = \frac{kD}{1 + hDe^{0k}} \qquad \Longrightarrow \qquad D = \frac{p_0}{k - hp_0}.$$

On peut donc conclure que la population évolue selon la fonction

$$p(t) = \begin{cases} 0 & \text{si } p_0 = 0, \\ \frac{k}{h} & \text{si } p_0 = \frac{k}{h}, \\ \frac{k}{\frac{k-hp_0}{p_0 e^{kt}} + h} & \text{sinon.} \end{cases}$$

Une simple étude de la fonction p montre que

- si $p_0 \in]0; k/h[$ alors $p'(t) > 0$ et $\lim_{t\to+\infty} p(t) = k/h$,
- si $p_0 \in]k/h; +\infty[$ alors $p'(t) < 0$ et $\lim_{t\to+\infty} p(t) = k/h$.



Exemple avec $k = 3$, $h = 1$ et différentes valeurs de p_0.

Solution approchée Supposons de ne pas vouloir/pouvoir calculer la fonction $p(t)$. Grâce à la méthode d'EULER on peut estimer le nombre d'indivus à différentes instantes 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 $\{p_i\}_i$ où chaque p_i est une approximation de $p(t_i)$. Si on utilise la méthode d'EULER, cette suite est ainsi construite :

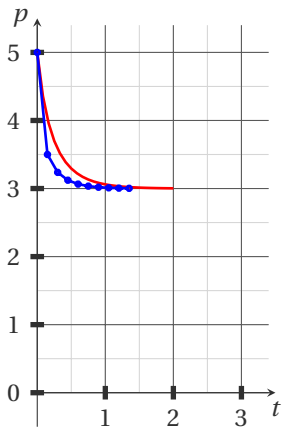
$$\begin{cases} p_{i+1} = p_i + \Delta t\ p_i(k - hp_i), \\ p_0 \text{ donné}, \end{cases}$$

qu'on peut réécrire comme

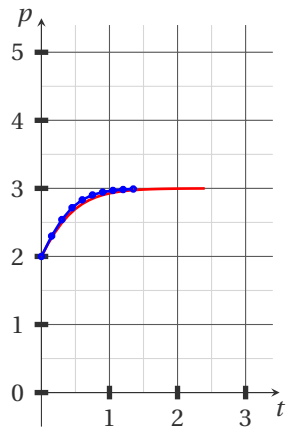
$$\begin{cases} p_{i+1} = (1 + k\Delta t - h\Delta t\, p_i)\, p_i, \\ p_0 \text{ donné}. \end{cases}$$

On veut appliquer cette méthode au cas de la figure précédente, i.e. avec $k = 3$, $h = 1$ et les valeurs initiales $p_0 = 5$ et $p_0 = 2$. Si on choisit comme pas temporelle $\Delta t = 0{,}15$, on obtient les figures suivantes :

t_i	$p(t_i)$	p_i	$p(t_i) - p_i$
0.00	5.000000	5.000000	0.000000
0.15	4.027123	3.500000	0.527123
0.30	3.582637	3.237500	0.345137
0.45	3.347079	3.122164	0.224915
0.60	3.212403	3.064952	0.147451
0.75	3.132046	3.035091	0.096956
0.90	3.082874	3.019115	0.063759
1.05	3.052319	3.010459	0.041861
1.20	3.033151	3.005736	0.027415
1.35	3.021054	3.003150	0.017904
1.50	3.013390	3.001731	0.011659
1.65	3.008524	3.000952	0.007573
1.80	3.005430	3.000523	0.004907



t_i	$p(t_i)$	p_i	$p(t_i) - p_i$
0.00	2.000000	2.000000	0.000000
0.15	2.274771	2.300000	−0.025229
0.30	2.493175	2.541500	−0.048325
0.45	2.655760	2.716292	−0.060532
0.60	2.770980	2.831887	−0.060907
0.75	2.849816	2.903298	−0.053483
0.90	2.902469	2.945411	−0.042942
1.05	2.937070	2.969529	−0.032459
1.20	2.959567	2.983102	−0.023535
1.35	2.974092	2.990663	−0.016571
1.50	2.983429	2.994852	−0.011423
1.65	2.989412	2.997164	−0.007752
1.80	2.993240	2.998439	−0.005199



CHAPITRE 6

Fonctions de plusieurs variables

- ⋆ Une *fonction* f de $\mathbb{R}^n$ et à valeurs réelles fait correspondre à tout point $\mathbf{x} \equiv (x_1, x_2, \ldots, x_n)$ 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 $\mathscr{D}_f \subset \mathbb{R}^n$ des points $\mathbf{x} \equiv (x_1, x_2, \ldots, x_n)$ qui ont une image par f.
- ⋆ L'*image* par f de $\mathscr{D}$ est l'ensemble $\mathrm{Im}_f(\mathscr{D}_f) = \{ r \in \mathbb{R} \mid r = f(\mathbf{x}),\ \mathbf{x} \in \mathbb{R}^n \} \subset \mathbb{R}$.
- ⋆ L'ensemble des points $S = \{ (\mathbf{x}, f(\mathbf{x})) \mid \mathbf{x} \in \mathscr{D}_f \}$ de $\mathbb{R}^{n+1}$ est la *surface représentative* de f ; c'est l'analogue 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

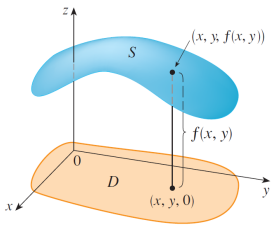
$$\mathscr{G}_f \equiv \{ (x, y, z = f(x, y)) \mid (x, y) \in \mathscr{D}_f \}$$

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

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

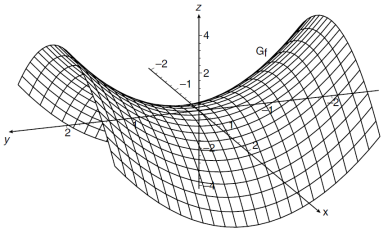
par l'altitude $z = f(x, y)$.



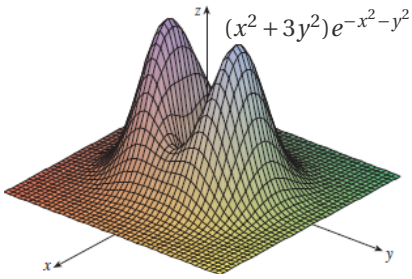
Les axes relatifs aux variables, x et y, sont conventionnellement situés dans un plan horizontal (le domaine $\mathscr{D}_f$ apparaît alors comme un sous-ensemble de ce plan), tandis que la dimension verticale est réservée aux valeurs de z. Ainsi, à tout $(x, y) \in \mathscr{D}_f$, dont l'image est $f(x, y) \in \mathbb{R}$, 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.

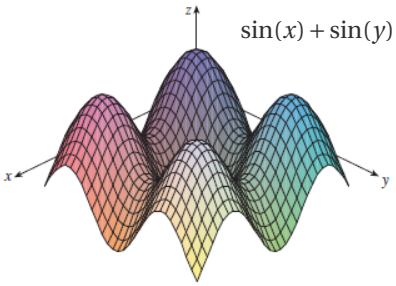
EXEMPLE

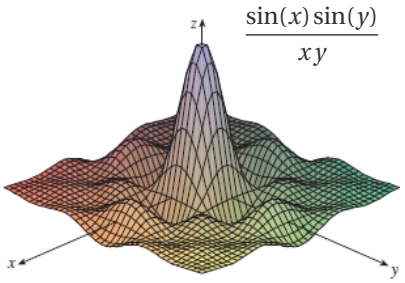
Le graphe de la fonction $f : \mathbb{R}^2 \to \mathbb{R}$ définie par $f(x, y) = x^2 - y^2$ est une surface de $\mathbb{R}^3$ qui a la forme d'une selle de cheval, comme l'indique la représentation en perspective de la figure ci-dessous.



Voici d'autres exemples :



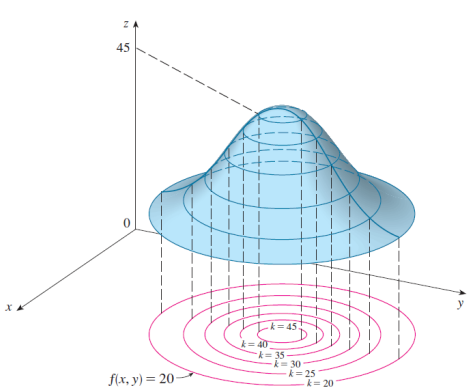


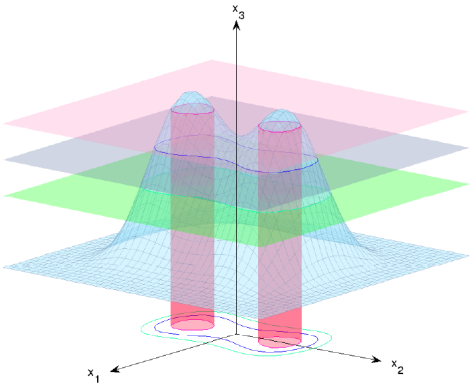


Si on considère des coupes horizontales on obtient, de façon générale, des courbes planes, dites *courbes ou lignes de niveau.*

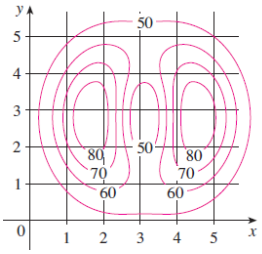
Définition 6.1 (Lignes de niveau)
Soit $k \in \mathbb{R}$ et f une fonction de $\mathbb{R}^2$ dans $\mathbb{R}$; la courbe de niveau k de la fonction f est la projection sur le plan d'équation $z = 0$ de l'intersection de la surface représentative de f avec le plan horizontal $z = k$, *i.e.* l'ensemble $\{(x, y) \in \mathscr{D} \mid f(x, y) = k\}$.

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.

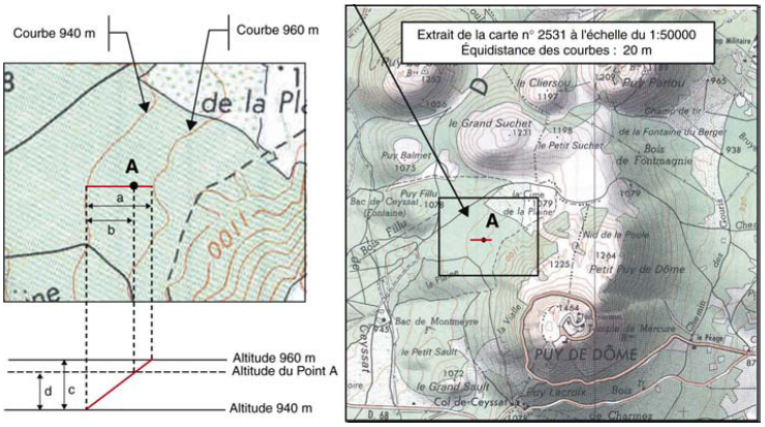




EXEMPLE
L'image ci-contre montre les courbes de niveaux d'une fonction f. On peut alors se faire une idée de l'allure de la fonction supposée continue. Par exemple $f(1;3) \approx 72$, $f(4;5) \approx 56$, soit $40 < f(3;3) < 50$ soit $50 < f(3;3) < 60$, etc.



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.



6.1. 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

$$\lim_{x\to x_0} \frac{f(x)-f(x_0)}{x-x_0}$$

ce qui équivaut, en posant $h = x - x_0$, à

$$\lim_{h\to 0} \frac{f(x_0+h)-f(x_0)}{h}.$$

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 $f\colon \mathbb{R}^2 \to \mathbb{R}$, 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 $f\colon \mathbb{R}^2 \to \mathbb{R}$ 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 6.2 (Dérivées partielles premières)
Soit f une fonction à valeurs réelles définie sur une partie ouverte $\mathscr{D}$ de $\mathbb{R}^2$. Soit $(x_0, y_0) \in \mathscr{D}$. 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

$$\frac{\partial f}{\partial x}(x_0,y_0) = \lim_{x\to x_0} \frac{f(x,y_0)-f(x_0,y_0)}{x-x_0} = \lim_{h\to 0} \frac{f(x_0+h,y_0)-f(x_0,y_0)}{h}$$

dérivée partielle de f par rapport à x au point (x_0, y_0)

$$\frac{\partial f}{\partial y}(x_0,y_0) = \lim_{y\to y_0} \frac{f(x_0,y)-f(x_0,y_0)}{y-y_0} = \lim_{k\to 0} \frac{f(x_0,y_0+k)-f(x_0,y_0)}{k}$$

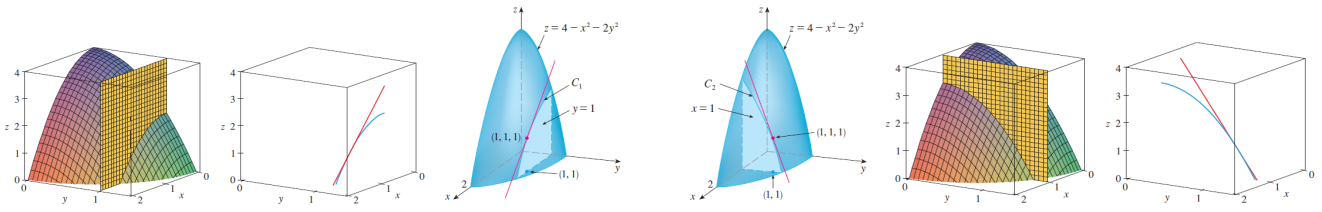
dérivée partielle de f par rapport à y au point (x_0, y_0)

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 $\partial_y f(1,1) = -4$.

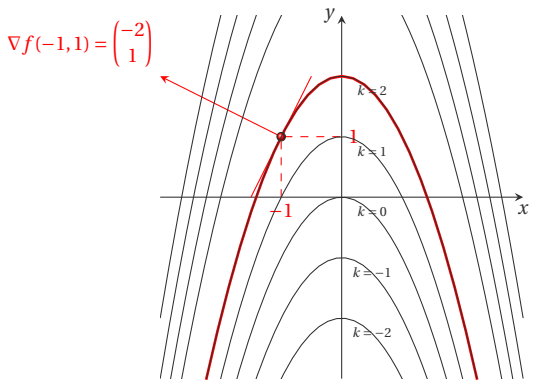


EXEMPLE
1. Soit la fonction $f(x,y) = 3x^2 + xy - 2y^2$. Alors $\mathscr{D}_f \equiv \mathbb{R}^2$, f est continue, $\partial_x f(x,y) = 6x + y$ (car y est considérée constante) et $\partial_y f(x,y) = x - 4y$ (car x est considérée constante).
2. Soit la fonction $f(x,y,z) = 5xz\ln(1+7y)$. Alors $\mathscr{D}_f \equiv \{(x,y,z) \mid y > -1/7\}$, 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 6.3 (Vecteur gradient)
Le gradient de la fonction $f : \mathbb{R}^n \to \mathbb{R}$ évalué au point $\hat{\mathbf{x}} = (\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n)$, noté $\nabla f(\hat{\mathbf{x}})$ ou encore **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 6.4 (Plan tangent)
Soit $\mathscr{D}$ une partie ouverte de $\mathbb{R}^n$ et soit $f : \mathscr{D} \to \mathbb{R}$ 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
$$L(x_0 + h, y_0 + k) = f(x_0,y_0) + h\partial_x f(x_0,y_0) + k\partial_y f(x_0,y_0).$$

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 $\hat{\mathbf{x}} = (-1,1)$ 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 $n-1$.

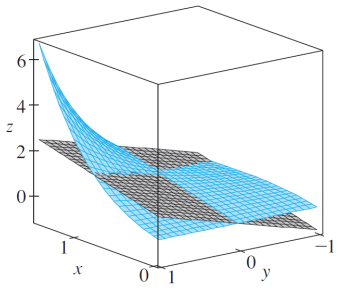
EXEMPLE
On peut calcule le plan tangent à la fonction $f : \mathbb{R}^2 \to \mathbb{R}$ définie par $f(x,y) = xe^{xy}$ en $(1,0)$ et utiliser sa linéarisation pour approcher $f(1.1, -0.1)$. On a

$$\begin{aligned} f(x,y) &= xe^{xy} & f(1,0) &= 1 \\ \partial_x f(x,y) &= e^{xy} + xye^{xy} & \partial_x f(1,0) &= 1 \\ \partial_y f(x,y) &= x^2 e^{xy} & \partial_y f(1,0) &= 1 \end{aligned}$$

Les trois fonctions f, $\partial_x f$ et $\partial_y f$ sont continues, donc f est différentiable. Sa linéarisation donne

$$f(x,y) \simeq f(1,0) + (x-1)\partial_x f(1,0) + (y-0)\partial_y f(1,0) = 1 + (x-1) + y = x + y,$$

autrement dit $xe^{xy} \simeq x + y$ lorsque $(x,y) \simeq (1,0)$, ainsi $f(1.1, -0.1) \simeq 1.1 - 0.1 = 1$. En effet, $f(1.1, -0.1) = 1.1e^{-0.11} \approx 0.98542$



6.2. 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 6.5 (Dérivées partielles d'ordre 2 pour une fonction de deux variables)
Soit la fonction $f : \mathscr{D} \subset \mathbb{R}^2 \to \mathbb{R}$ où $\mathscr{D}_f$ 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 :

$$\frac{\partial^2 f}{\partial x^2}(x_0,y_0) = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right)(x_0,y_0) \qquad \text{(notée aussi } \partial_{xx} f(x_0,y_0)),$$

$$\frac{\partial^2 f}{\partial x\,\partial y}(x_0,y_0) = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right)(x_0,y_0) \qquad \text{(notée aussi } \partial_{xy} f(x_0,y_0)),$$

$$\frac{\partial^2 f}{\partial y\,\partial x}(x_0,y_0)=\frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right)(x_0,y_0) \qquad \text{(notée aussi } \partial_{yx}f(x_0,y_0)\text{)},$$
$$\frac{\partial^2 f}{\partial y^2}(x_0,y_0)=\frac{\partial}{\partial y}\left(\frac{\partial f}{\partial y}\right)(x_0,y_0) \qquad \text{(notée aussi } \partial_{yy}f(x_0,y_0)\text{)}.$$

Les dérivées partielles d'ordre supérieur à 2 se définissent par récurrence de façon analogue. Soit la fonction $f:\mathbb{R}^n\to\mathbb{R}$; 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 6.6 (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 6.7 (Matrice hessienne)
Soit la fonction $f:\mathscr{D}\subset\mathbb{R}^2\to\mathbb{R}$ où $\mathscr{D}_f$ est un ouvert de $\mathbb{R}^2$. La matrice hessienne de f en (x_0,y_0) est la matrice de taille 2×2 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 $n>2$.

EXEMPLE
Les dérivées premières et secondes de la fonction $f(x,y)=-2x^2+3xy^2-y^3$ sont

$$\partial_x f(x,y)=-4x+3y^2,\qquad \partial_y f(x,y)=6xy-3y^2,$$
$$\partial_{xx}f(x,y)=-4,\qquad \partial_{xy}f(x,y)=6y,\qquad \partial_{yx}f(x,y)=6y,\qquad \partial_{yy}f(x,y)=6x-6y.$$

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.

6.3. Optimisation (dans un ouvert et sans contraintes)

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 6.8
Soit f une fonction de $\mathscr{D}\subset\mathbb{R}^n$ dans $\mathbb{R}$. On dit que

- f est bornée dans $\mathscr{D}$ s'il existe un nombre réel $M\geq 0$ tel que
$$\forall\mathbf{x}\in\mathscr{D},\ |f(\mathbf{x})|\leq M;$$
- f admet un maximum (resp. minimum) *global* (ou absolu) en $\mathbf{x}_0\in\mathscr{D}$ si
$$\forall\mathbf{x}\in\mathscr{D},\ f(\mathbf{x})\leq f(\mathbf{x}_0)\ \text{(resp. } f(\mathbf{x})\geq f(\mathbf{x}_0)\text{)};$$
- f admet un maximum (resp. minimum) *local* (ou relatif) en $\mathbf{x}_0\in\mathscr{D}$ s'il existe une boule de rayon non nul $\mathscr{B}(\mathbf{x}_0,r)$ telle que
$$\forall\mathbf{x}\in\mathscr{D}\cap\mathscr{B}(\mathbf{x}_0,r),\ f(\mathbf{x})\leq f(\mathbf{x}_0)\ \text{(resp. } f(\mathbf{x})\geq f(\mathbf{x}_0)\text{)}.$$

Théorème 6.9 (de FERMAT : condition nécessaire du premier ordre)
Soit $\mathscr{D}$ un sous-ensemble ouvert de $\mathbb{R}^n$, $\mathbf{x}_0$ un point contenu dans $\mathscr{D}$ et $f:\mathscr{D}\to\mathbb{R}$ une fonction de classe $\mathscr{C}^1$ en ce point. Si f présente un extrémum local alors
$$\nabla f(\mathbf{x}_0)=\mathbf{0}.$$

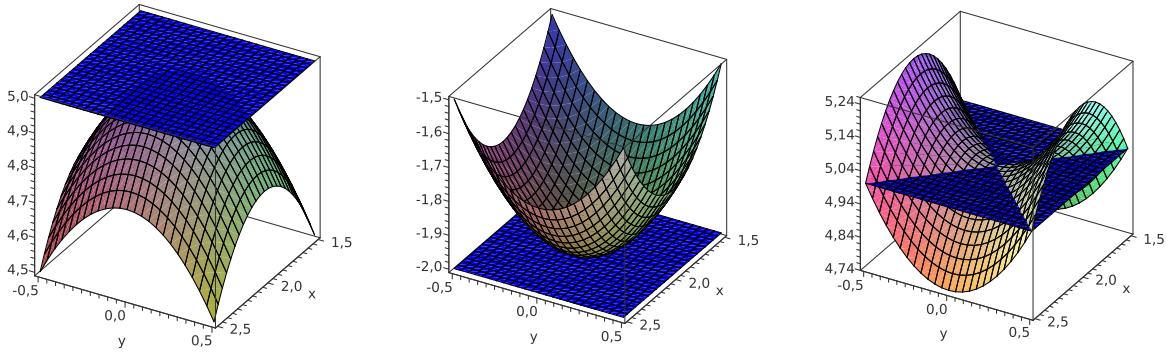
Définition 6.10 (Point stationnaire ou critique)
À l'instar des fonctions d'une variable réelle, un point $\mathbf{x}_0$ vérifiant $\nabla f(\mathbf{x}_0) = \mathbf{0}$ est appelé *point stationnaire* ou *point critique* de f.

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 extremum 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 $\mathscr{D} = \{(x,y)\in\mathbb{R}^2 \mid x^2+y^2<1\}$. Le seul candidat extremum 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

$$f(x,y) = x^2+y^2 \geq 0 = f(0,0) \qquad \forall (x,y)\in\mathscr{D}_f.$$

En revanche, la fonction n'admet aucun maximum.

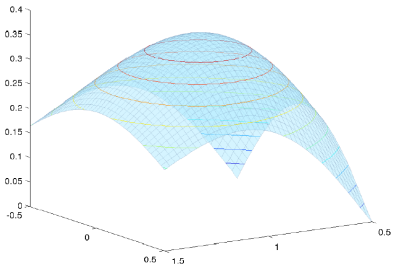
Théorème 6.11 (Condition suffisante d'extrémum local dans un ouvert (cas de 2 variables))
Soit f une fonction de classe $\mathscr{C}^2$ sur un ouvert $\mathscr{D}\subset\mathbb{R}^2$ et (x_0,y_0) un point stationnaire ; posons

$$\text{dét}(H_f(x_0,y_0)) \equiv \partial_{xx}f(x_0,y_0)\cdot\partial_{yy}f(x_0,y_0) - \big(\partial_{xy}f(x_0,y_0)\big)^2,$$

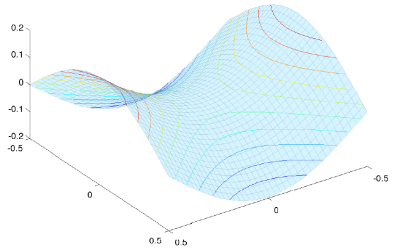
le déterminant de la matrice hessienne de f évalué en (x_0,y_0).

- Si $\text{dét}(H_f(x_0,y_0))>0$, 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 $\text{dét}(H_f(x_0,y_0))<0$, alors f présente un point-selle (ou point-col) en (x_0,y_0) ; ce n'est pas un extrémum ;
- si $\text{dét}(H_f(x_0,y_0))=0$, 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 tableaux suivant :



(a) Point de maximum



(b) Point de selle

$\det(H_f(x_0,y_0))$	$\partial_{xx}f(x_0,y_0)$	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 extremum 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 $\left(-\frac{1}{3},-\frac{1}{3}\right)$ et $(1,1)$.

Classification La matrice hessienne de f en un point $(x,y)\in\mathbb{R}^2$ est

$$H_f(x,y)=\begin{pmatrix}\partial_{xx}f(x,y) & \partial_{xy}f(x,y)\\ \partial_{yx}f(x,y) & \partial_{yy}f(x,y)\end{pmatrix}=\begin{pmatrix}2 & -2\\ -2 & 6y\end{pmatrix}.$$

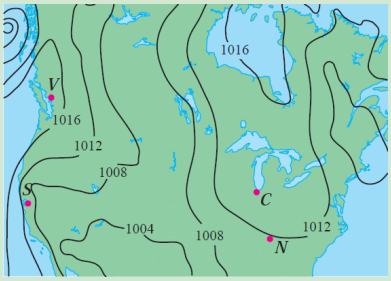
Comme $\det\left(H_f\left(-\frac{1}{3},-\frac{1}{3}\right)\right)<0$ et $D(1,1)>0$, alors $\left(-\frac{1}{3},-\frac{1}{3}\right)$ est un point-selle et f admet en $(1,1)$ un minimum local de valeur $f(1,1)=-1$. Ce minimum n'est cependant pas global puisque, par exemple, $f(0,-2)=-6<f(1,1)=-1$.

6.4. Exercices

Exercice 6.1

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 6.2

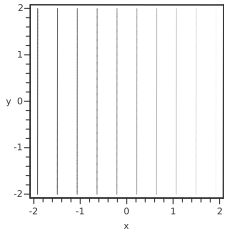
Déterminer les courbes de niveau des fonctions suivantes :

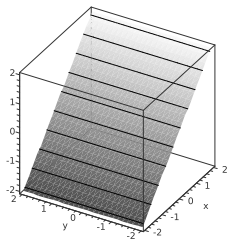
$$f(x,y)=x, \qquad f(x,y)=y+1, \qquad f(x,y)=x+y-1, \qquad f(x,y)=e^{y-x^2}, \qquad f(x,y)=y-\cos(x).$$

Esquissez ensuite leurs graphes (le graphe peut être vu comme un empilement de courbes de niveau qui forment une surface dans $\mathbb{R}^3$).

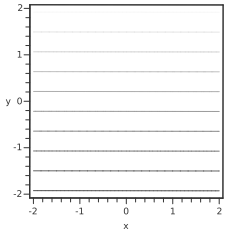
Correction

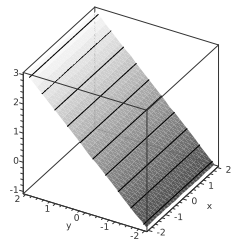
⋆ $f(x,y)=x$:
$f(x,y)=\kappa$ ssi $x=\kappa$, les courbes de niveau sont des droites verticales et la surface représentative de f est un plan.



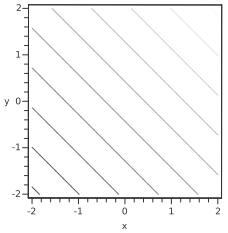


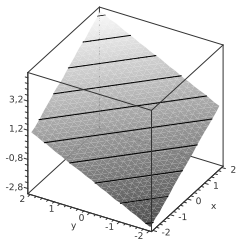
⋆ $f(x,y)=y+1$:
$f(x,y)=\kappa$ ssi $y=\kappa-1$, les courbes de niveau sont des droites horizontales et la surface représentative de f est un plan.



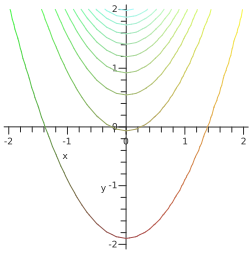
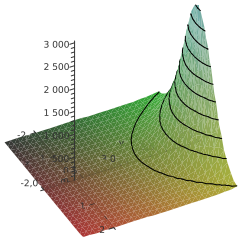


⋆ $f(x,y)=x+y-1$:
$f(x,y)=\kappa$ ssi $y=-x+(\kappa+1)$, les courbes de niveau sont des droites de pente -1 et la surface représentative de f est un plan.

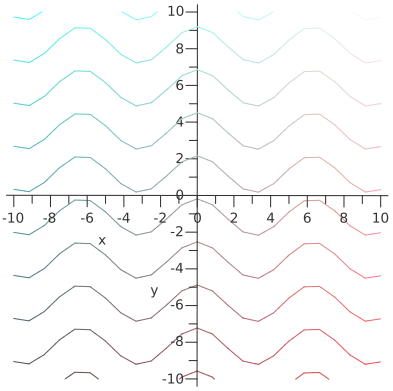
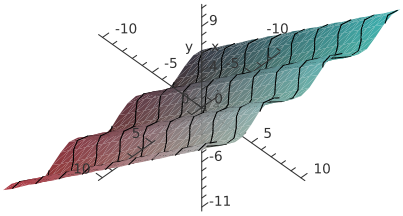




⋆ $f(x,y)=e^{y-x^2}$:
$f(x,y)=\kappa$ ssi $y=x^2+\ln(\kappa)$, les courbes de niveau sont des paraboles. On observe notamment la croissance exponentielle marquée lorsque les valeurs prises par y sont grandes et celles prises par $|x|$ sont petites.

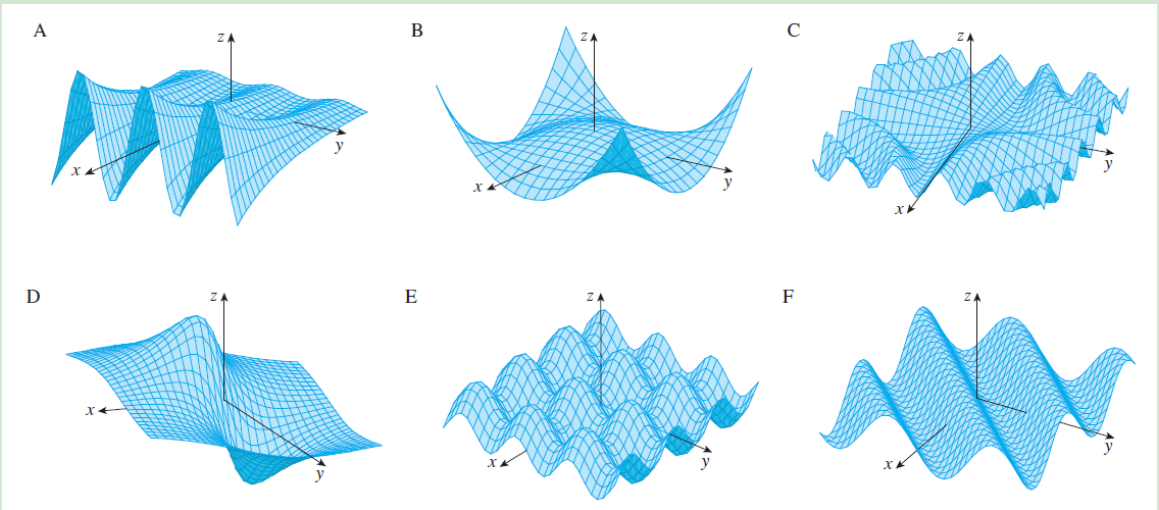
⋆ $f(x,y) = y - \cos(x)$:
$f(x,y) = \kappa$ ssi $y = \cos(x) + \kappa$

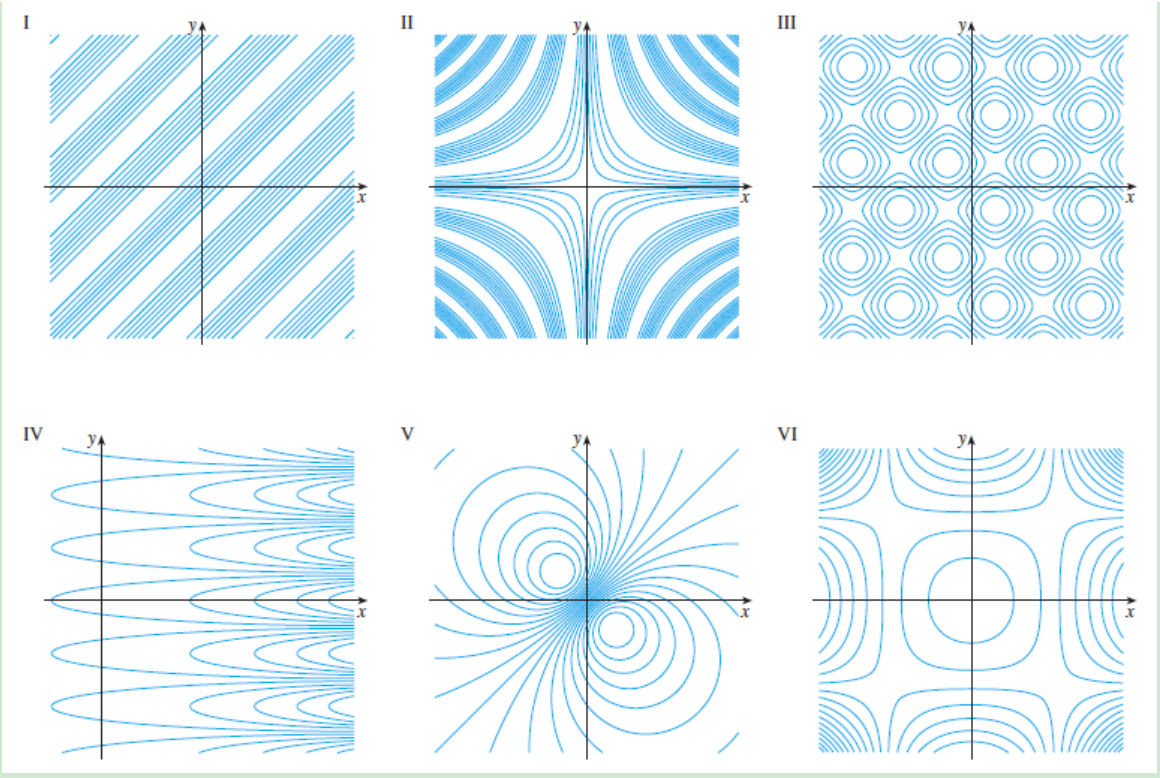



Exercice 6.3

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. $f(x,y) = \dfrac{x-y}{1+x^2+y^2}$
5. $f(x,y) = e^x \cos(y)$
6. $f(x,y) = \sin(x) - \sin(y)$

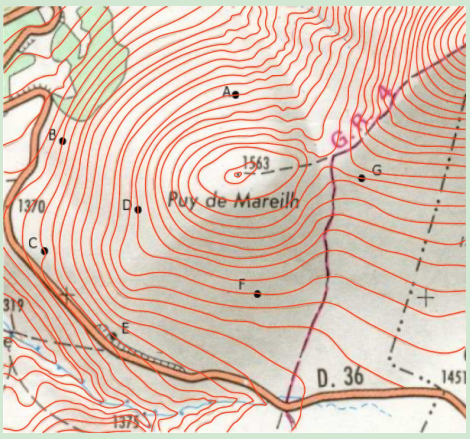




Correction

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 6.4 (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 6.5

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. $f(x,y) = \frac{x}{y}$
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. $z(x,y) = (2x+3y)^{10}$
9. $f(x,y) = \frac{ax+by}{cx+dy}$
10. $F(x,y) = \int_y^x \cos(e^t)\,\mathrm{d}t$

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 $\partial_y f(x,y) = \ln(x)x^y$ ⚠ $x^y = e^{y\ln(x)}$ donc $x > 0$
6. $\partial_x f(x,y,z) = \cos(xz) - xz\sin(xz)$, $\partial_y f(x,y,z) = \dfrac{-2\sin(y+z)\cos(y+z)}{2-\sin^2(y+z)}$ et $\partial_z f(x,y,z) = -x^2\sin(xz) + \dfrac{-2\sin(y+z)\cos(y+z)}{2-\sin^2(y+z)}$
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 $\partial_y F(x,y) = -\cos(e^y)$

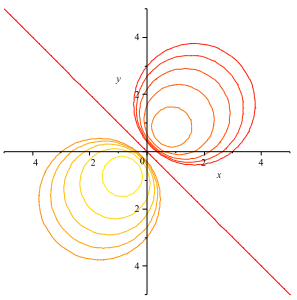
Exercice 6.6

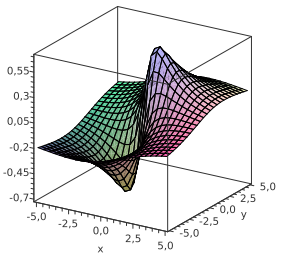
Soit $f : \mathbb{R}^2 \to \mathbb{R}$ la fonction définie par $f(x,y) = \dfrac{x+y}{1+x^2+y^2}$.

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 $\left(\frac{1}{2k}, \frac{1}{2k}\right)$ et rayon $\sqrt{\frac{1}{2k^2} - 1}$.





2. Les deux dérivées premières partielles de f sont

$$\partial_x f(x,y) = \frac{1-x^2-2xy+y^2}{(1+x^2+y^2)^2}, \qquad \partial_y f(x,y) = \frac{1+x^2-2xy-y^2}{(1+x^2+y^2)^2}.$$

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 6.7

Soit $f\colon \mathbb{R}^2 \to \mathbb{R}$ une fonction de classe $\mathscr{C}^2(\mathbb{R}^2)$ 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 6.8

On suppose que $(1,1)$ est un point critique d'une fonction f dont les dérivées secondes sont continues. Dans chaque cas, que peut-on dire au sujet de f ?

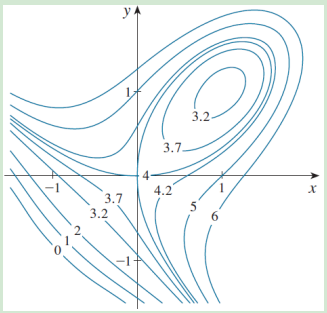
1. $\partial_{xx} f(1,1)=4, \partial_{xy} f(1,1)=1, \partial_{yy} f(1,1)=2$;
2. $\partial_{xx} f(1,1)=4, \partial_{xy} f(1,1)=3, \partial_{yy} f(1,1)=2$.

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 6.9

À partir de la carte des courbes de niveau de la figure ci-contre, localiser les points critiques de $f\colon \mathbb{R}^2 \to \mathbb{R}$ 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

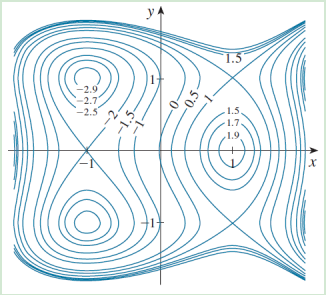
Dans la figure, le point $(1,1)$ est entouré par des courbes de niveau qui sont de forme ovale et qui indiquent que si nous nous éloignons du point dans n'importe quelle direction les valeurs de f augmentent. Ainsi on pourrait s'attendre à un minimum local en ou à proximité de $(1,1)$.
Les courbes de 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 6.10

À partir de la carte des courbes de niveau de la figure ci-contre, localiser les points critiques de $f : \mathbb{R}^2 \to \mathbb{R}$ 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 6.11

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 :

$$\partial_{xx} f(x,y) = -4 < 0, \qquad \partial_{yy} f(x,y) = -2, \qquad \partial_{xy} f(x,y) = 2,$$

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 6.12

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 $f\colon \mathbb{R}^2 \to \mathbb{R}$ définie par $f(x,y) = 4y^2e^x - 2y^4 - e^{4x}$. Montrer que cette fonctions admet deux maxima mais aucun autre point critique.

Correction

⋆ f est de classe $\mathscr{C}^2$ dans son domaine de définition, l'ouvert $\mathbb{R}^2$.

⋆ Recherche de points critiques :

$$\nabla f(x,y) = \begin{pmatrix}0\\0\end{pmatrix} \iff \begin{cases} 4y^2e^x - 4e^{4x} = 0 \\ 8ye^x - 8y^3 = 0 \end{cases} \iff \begin{cases} 4e^x(y^2 - e^{3x}) = 0 \\ 8y(e^x - y^2) = 0 \end{cases} \nearrow\searrow \begin{cases} y = 0 \\ -4e^{4x} = 0 \end{cases} \text{ ou } \begin{cases} e^x = y^2 \\ y^2 = e^{3x} = y^6 \end{cases} \iff (x,y) = (0,\pm 1).$$

On a deux points critiques : $(0,1)$ et $(0,-1)$.

⋆ Nature des points critiques :

$$H_f(x,y) = \begin{pmatrix} 4y^2e^x - 16e^{4x} & 8ye^x \\ 8ye^x & 8e^x - 24y^2 \end{pmatrix}, \qquad \text{dét}(H_f(x,y)) = 32e^x\left((e^x - 3y^2)(y^2 - 4e^{3x}) - 2y^2e^x\right).$$

$\det\big(H_f(0,\pm 1)\big) = 128 > 0$ et $\partial_{xx}f(0,\pm 1,) = -12 < 0$ donc les points $(0,\pm 1)$ sont des maxima.

Exercice 6.13

Déterminer et établir la nature des points critiques des fonction $f\colon \mathbb{R}^2 \to \mathbb{R}$ 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. $f(x,y) = \frac{x^2y}{2} + x^2 + \frac{y^3}{3} - 4y$
13. $f(x,y) = \frac{x^2y}{2} - x^2 + \frac{y^3}{3} - 4y$
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. $f(x,y) = x^3y^2(6-x-y)$
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 $\mathscr{C}^2$ 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 : $\left(\frac{1}{3}, -\frac{2}{3}\right)$.

⋆ Nature du point critique :

$$H_f(x,y) = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \qquad \text{dét}(H_f(x,y)) = 3.$$

$\text{dét}\left(H_f\left(\frac{1}{3}, -\frac{2}{3}\right)\right) > 0$ et $\partial_{xx}f\left(\frac{1}{3}, -\frac{2}{3}\right) > 0$ donc $\left(\frac{1}{3}, -\frac{2}{3}\right)$ est un minimum.

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

⋆ f est de classe $\mathscr{C}^2$ 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 :

$$H_f(x,y)=\begin{pmatrix}-2 & 1\\ 1 & -2\end{pmatrix}, \qquad \text{dét}(H_f(x,y))=3.$$

$\text{dét}(H_f(-2,-2))>0$ 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 $\mathscr{C}^2$ 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 :

$$H_f(x,y)=\begin{pmatrix}-2y & -2x+2y\\ -2x+2y & 2x\end{pmatrix}, \qquad \text{dét}(H_f(x,y))=-4xy-4(y-x)^2.$$

$\text{dét}(H_f(-1,-1))<0$ donc $(-1,-1)$ est un point-selle ;
$\text{dét}(H_f(1,1))<0$ donc $(1,1)$ est un point-selle.

4. $f(x,y)=y^3+3x^2y-6x^2-6y^2+2$

⋆ f est de classe $\mathscr{C}^2$ 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 :

$$H_f(x,y)=\begin{pmatrix}6y-12 & 6x\\ 6x & 6y-12\end{pmatrix}, \qquad \text{dét}(H_f(x,y))=(6y-12)^2-36x^2=36\Big((y-2)^2-x^2\Big).$$

$\text{dét}(H_f(0,0))>0$ et $\partial_{xx}f(0,0)<0$ donc $(0,0)$ est un maximum ;
$\text{dét}(H_f(0,4))>0$ et $\partial_{xx}f(0,4)>0$ donc $(0,4)$ est un minimum ;
$\text{dét}(H_f(2,2))<0$ donc $(2,2)$ est un point-selle ;
$\text{dét}(H_f(-2,2))<0$ donc $(-2,2)$ est un point-selle.

5. $f(x,y)=x^3+y^3-3xy+3$.

⋆ f est de classe $\mathscr{C}^2$ 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 :

$$H_f(x,y)=\begin{pmatrix}6x & -3\\ -3 & 6y\end{pmatrix}, \qquad \text{dét}(H_f(x,y))=36xy-9.$$

$\text{dét}(H_f(0,0))=-9<0$ donc $(0,0)$ est un point-selle ;
$\text{dét}(H_f(1,1))=27>0$ 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 $\mathscr{C}^2$ dans son domaine de définition, l'ouvert $\mathbb{R}^2$.
   - ⋆ Recherche de points critiques :

$$\nabla f(x,y) = \begin{pmatrix}0\\0\end{pmatrix} \iff \begin{cases} y-2xy-y^2=0 \\ x-x^2-2xy=0 \end{cases} \iff (x,y) \in \left\{ (0,0),(1,0),(0,1),\left(\frac{1}{3},\frac{1}{3}\right)\right\}.$$

   On a quatre points critiques : $(0,0)$, $(1,0)$, $(0,1)$ et $\left(\frac{1}{3},\frac{1}{3}\right)$.
   - ⋆ Nature des points critiques :

$$H_f(x,y) = \begin{pmatrix} -2y & 1-2x-2y \\ 1-2x-2y & -2x \end{pmatrix}, \qquad \text{dét}(H_f(x,y)) = 4xy-(1-2x-2y)^2$$

   $\text{dét}(H_f(0,0)) < 0$ donc $(0,0)$ est un point-selle ;
   $\text{dét}(H_f(1,0)) < 0$ donc $(1,0)$ est un point-selle ;
   $\text{dét}(H_f(0,1)) < 0$ donc $(0,1)$ est un point-selle ;
   $\text{dét}\left(H_f\left(\frac{1}{3},\frac{1}{3}\right)\right) > 0$ et $\partial_{xx} f\left(\frac{1}{3},\frac{1}{3}\right) < 0$ donc $\left(\frac{1}{3},\frac{1}{3}\right) < 0$ est un maximum.
7. $f(x,y) = x^3 - 12xy + 8y^3$
   - ⋆ f est de classe $\mathscr{C}^2$ 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 :

$$H_f(x,y) = \begin{pmatrix} 6x & -12 \\ -12 & 48y \end{pmatrix}, \qquad \text{dét}(H_f(x,y)) = 144(2xy-1)$$

   $\text{dét}(H_f(0,0)) < 0$ donc $(0,0)$ est un point-selle ;
   $\text{dét}(H_f(2,1)) > 0$ 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 $\mathscr{C}^2$ 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 :

$$H_f(x,y) = \begin{pmatrix} \frac{2}{x^3} & 1 \\ 1 & \frac{2}{y^3} \end{pmatrix}, \qquad \text{dét}(H_f(x,y)) = \frac{4}{(xy)^3} - 1$$

   $\text{dét}(H_f(1,1)) > 0$ 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 $\mathscr{C}^2$ dans son domaine de définition, l'ouvert $\mathbb{R}^2$.
   - ⋆ Recherche de points critiques :

$$\nabla f(x,y) = \begin{pmatrix}0\\0\end{pmatrix} \iff \begin{cases} e^x\cos(y)=0 \\ -e^x\sin(y)=0 \end{cases} \iff \nexists (x,y) \in \mathbb{R}^2.$$

   Cette fonction n'admet aucun point critique.
10. $f(x,y) = y\cos(x)$
   - ⋆ f est de classe $\mathscr{C}^2$ 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 :

$$H_f(x,y)=\begin{pmatrix}-y\cos(x)&-\sin(x)\\-\sin(x)&0\end{pmatrix},\qquad\text{dét}(H_f(x,y))=-\sin^2(x)$$

$\text{dét}\left(H_f\left(\frac{\pi}{2}+\kappa\pi,0\right)\right)<0$ donc ils sont tous des points-selle.

11. $f(x,y)=y^2+xy\ln(x)$

⋆ f est de classe $\mathscr{C}^2$ dans son domaine de définition, l'ouvert $\mathscr{D}=\left\{(x,y)\in\mathbb{R}^2\;\middle|\;x>0\right\}$.

⋆ 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 $\left(\frac{1}{e},\frac{1}{2e}\right)$.

⋆ Nature des points critiques :

$$H_f(x,y)=\begin{pmatrix}\frac{y}{x}&1+\ln(x)\\1+\ln(x)&2\end{pmatrix},\qquad\text{dét}(H_f(x,y))=2\frac{y}{x}-(1+\ln(x))^2$$

$\text{dét}(H_f(1,0))<0$ donc $(1,0)$ est un point-selle ;
$\text{dét}\left(H_f\left(\frac{1}{e},\frac{1}{2e}\right)\right)>0$ et $\partial_{xx}f\left(\frac{1}{e},\frac{1}{2e}\right)>0$ donc $\left(\frac{1}{e},\frac{1}{2e}\right)$ est un minimum.

12. $f(x,y)=\frac{x^2y}{2}+x^2+\frac{y^3}{3}-4y$

⋆ f est de classe $\mathscr{C}^2$ 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 :

$$H_f(x,y)=\begin{pmatrix}y+2&x\\x&2y\end{pmatrix},\qquad\text{dét}(H_f(x,y))=2y(y+2)-x$$

$\text{dét}(H_f(0,2))>0$ et $\partial_{xx}f(0,2)=4>0$ donc $(0,2)$ est un minimum pour f ;
comme $\text{dét}(H_f(0,-2))=0$, on ne peut pas conclure en utilisant la matrice hessienne (l'étude du signe de la distance dans ce cas est trop compliquée).

13. $f(x,y)=\frac{x^2y}{2}-x^2+\frac{y^3}{3}-4y$

⋆ f est de classe $\mathscr{C}^2$ 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 :

$$H_f(x,y)=\begin{pmatrix}y-2&x\\x&2y\end{pmatrix},\qquad\text{dét}(H_f(x,y))=2y(y-2)-x$$

dét$(H_f(0,-2)) > 0$ et $\partial_{xx} f(0,-2) < 0$ donc $(0,-2)$ est un maximum pour f ;
comme dét$(H_f(0,2)) = 0$, 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 $\mathscr{C}^2$ 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 :

$$H_f(x,y) = \begin{pmatrix} 2x & y \\ y & x+2 \end{pmatrix}, \qquad \text{dét}(H_f(x,y)) = 2x(x+2) - y$$

dét$(H_f(2,0)) > 0$ et $\partial_{xx} f(2,0) = 4 > 0$ donc $(2,0)$ est un minimum pour f ;
comme dét$(H_f(-2,0)) = 0$, 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 $\mathscr{C}^2$ 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 :

$$\begin{aligned} \partial_{xx} f(x,y) &= 2e^{(-x^2-y^2)}(1 - 5x^2 + y^2 + 2x^4 - 2x^2y^2), \\ \partial_{xy} f(x,y) &= 4xy(x^2 - y^2)e^{(-x^2-y^2)}, \\ \partial_{yy} f(x,y) &= 2e^{(-x^2-y^2)}(-1 - x^2 + 5y^2 + 2x^2y^2 - 2y^4). \end{aligned}$$

$$H_f(x,y) = \begin{pmatrix} \partial_{xx} f(x,y) & \partial_{xy} f(x,y) \\ \partial_{xy} f(x,y) & \partial_{yy} f(x,y) \end{pmatrix}, \qquad \text{dét}(H_f(x,y)) = \partial_{xx} f(x,y)\partial_{yy} f(x,y) - (\partial_{xy} f(x,y))^2$$

On a alors

(x_0,y_0)	$\partial_{xx} f(x_0,y_0)$	$\partial_{xy} f(x_0,y_0)$	$\partial_{yy} f(x_0,y_0)$	dét$(H_f(x_0,y_0))$	
$(0,0)$	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 $\mathscr{C}^2$ 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 :

$$\partial_{xx}f(x,y) = -2e^{(-x^2-y^2)}(1-5x^2+y^2+2x^4-2x^2y^2),$$
$$\partial_{xy}f(x,y) = -4xy(x^2-y^2)e^{(x^2-y^2)},$$
$$\partial_{yy}f(x,y) = -2e^{(-x^2-y^2)}(-1-x^2+5y^2+2x^2y^2-2y^4).$$

$$H_f(x,y) = \begin{pmatrix} \partial_{xx}f(x,y) & \partial_{xy}f(x,y) \\ \partial_{xy}f(x,y) & \partial_{yy}f(x,y) \end{pmatrix}, \qquad \text{dét}(H_f(x,y)) = \partial_{xx}f(x,y)\partial_{yy}f(x,y) - (\partial_{xy}f(x,y))^2$$

On a alors

(x_0,y_0)	$\partial_{xx}f(x_0,y_0)$	$\partial_{xy}f(x_0,y_0)$	$\partial_{yy}f(x_0,y_0)$	$\text{dét}(H_f(x_0,y_0))$	
$(0,0)$	-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 $\mathscr{C}^2$ 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 :

$$\nabla f(x,y) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \begin{cases} 4(x^3-x+y) = 0 \\ 4(y^3+x-y) = 0 \end{cases} \iff (x,y) \in \Big\{ (0,0), (\sqrt{2},-\sqrt{2}), (-\sqrt{2},\sqrt{2}) \Big\}.$$

On a 3 points critiques : [1] $(0,0)$, $(\sqrt{2},-\sqrt{2})$ et $(-\sqrt{2},\sqrt{2})$ (on note que $f(x,y) = f(-x,-y)$).

⋆ Nature des points critiques :

$$H_f(x,y) = \begin{pmatrix} 12x^2-4 & 4 \\ 4 & 12y^2-4 \end{pmatrix}, \qquad \text{dét}(H_f(x,y)) = 16\Big((3x^2-1)(3y^2-1)-1\Big).$$

$\text{dét}(H_f(\sqrt{2},-\sqrt{2})) = 384 > 0$ et $\partial_{xx}f(\sqrt{2},-\sqrt{2}) = 20 > 0$ donc $(\sqrt{2},-\sqrt{2})$ est un minimum pour f ;
$\text{dét}(H_f(-\sqrt{2},\sqrt{2})) = 384 > 0$ et $\partial_{xx}f(-\sqrt{2},\sqrt{2}) = 20 > 0$ donc $(-\sqrt{2},\sqrt{2})$ est un minimum pour f ;
comme $\text{dét}(H_f(0,0)) = 0$, 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

$$f(x,y) = (x^2-2)^2+(y^2-2)^2+2(x+y)^2-8 \geq 8 \qquad \forall (x,y) \in \mathbb{R}^2.$$

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 $\mathscr{C}^2$ 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 :

$$\nabla f(x,y) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \begin{cases} 4(x^3-2x+2y) = 0 \\ 4(y^3+2x-2y) = 0 \end{cases} \iff (x,y) \in \{ (0,0), (2,-2), (-2,2) \}.$$

1. $\begin{cases} 4x^3-4x+4y = 0 \\ 4y^3+4x-4y = 0 \end{cases} \implies \begin{cases} x^3+y^3 = 0 \\ y^3+x-y = 0 \end{cases} \implies \begin{cases} x = -y \\ (y^2-2)y = 0 \end{cases}$

On a 3 points critiques : [2] $(0,0)$, $(2,-2)$ et $(-2,2)$ (on note que $f(x,y) = f(-x,-y)$).

- ⋆ Nature des points critiques :

$$H_f(x,y) = \begin{pmatrix} 12x^2-8 & 8 \\ 8 & 12y^2-8 \end{pmatrix}, \qquad \text{dét}(H_f(x,y)) = 48\Big(3x^2y^2 - 2(x^2+y^2)\Big).$$

- ⋆ $\text{dét}(H_f(2,-2)) = 1536 > 0$ et $\partial_{xx} f(2,-2) = 40 > 0$ donc $(2,-2)$ est un minimum local pour f ;
- ⋆ $\text{dét}(H_f(-2,2)) = 1536 > 0$ et $\partial_{xx} f(-2,2) = 40 > 0$ donc $(-2,2)$ est un minimum local pour f ;
- ⋆ comme $\text{dét}(H_f(0,0)) = 0$, 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

$$f(x,y) = (x^2-4)^2 + (y^2-4)^2 + 4(x+y)^2 - 32 \geq -32 \qquad \forall (x,y) \in \mathbb{R}^2.$$

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 Δf 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 $\mathscr{C}^2$ 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 :

$$H_f(x,y) = \begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix}, \qquad \text{dét}(H_f(x,y)) = 8.$$

$\text{dét}(H_f(1,0)) = 8 > 0$ 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 $\mathscr{C}^2$ dans son domaine de définition, l'ouvert $\mathbb{R}^2$.

2. $\begin{cases} 4x^3 - 8x + 8y = 0 \\ 4y^3 + 8x - 8y = 0 \end{cases} \implies \begin{cases} x^3 - 2(x-y) = 0 \\ y^3 + 2(x-y) = 0 \end{cases} \implies \begin{cases} x^3 + y^3 = 0 \\ y^3 + 2x - 2y = 0 \end{cases} \implies \begin{cases} x = -y \\ (y^2-4)y = 0 \end{cases} \implies \begin{cases} x = -y \\ y = 0 \text{ ou } y = 2 \text{ ou } y = -2 \end{cases}$

⋆ 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 :

$$H_f(x,y)=\begin{pmatrix}2&1\\1&4\end{pmatrix},\qquad \text{dét}(H_f(x,y))=7.$$

$\text{dét}(H_f(1,0))=8>0$ 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 $\mathscr{C}^2$ 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 $t\in_R$ sont des points critiques ainsi que le point $(3,2)$.

⋆ Nature des points critiques :

$$\begin{aligned}\partial_{xx}f(x,y)&=6xy^2(6-x-y)-6x^2y^2,\\\partial_{xy}f(x,y)&=6x^2y(6-x-y)-3x^2y^2-2x^3y,\\\partial_{yy}f(x,y)&=2x^3(6-x-y)-4x^3y.\end{aligned}$$

$$H_f(x,y)=\begin{pmatrix}\partial_{xx}f(x,y)&\partial_{xy}f(x,y)\\\partial_{xy}f(x,y)&\partial_{yy}f(x,y)\end{pmatrix},\qquad \text{dét}(H_f(x,y))=\partial_{xx}f(x,y)\partial_{yy}f(x,y)-(\partial_{xy}f(x,y))^2$$

$\text{dét}(H_f(3,2))>0$ et $\partial_{xx}f(3,2)<0$ donc $(3,2)$ est un maximum pour f.
$\text{dét}(H_f(t,0))=0$ 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.
$\text{dét}(H_f(0,t))=0$ 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 $\mathscr{C}^2$ 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 :

$$\partial_{xx}f(x,y)=e^{x-y}(x^2-2y^2+4x+2),\qquad \partial_{xy}f(x,y)=e^{x-y}(-x^2+2y^2-2x-4y),\qquad \partial_{yy}f(x,y)=e^{x-y}(x^2-2y^2+8y-4);$$

$$H_f(x,y)=\begin{pmatrix}\partial_{xx}f(x,y)&\partial_{xy}f(x,y)\\\partial_{xy}f(x,y)&\partial_{yy}f(x,y)\end{pmatrix},\qquad \text{dét}(H_f(x,y))=\partial_{xx}f(x,y)\partial_{yy}f(x,y)-(\partial_{xy}f(x,y))^2.$$

On en déduit que

(x_0,y_0)	$\partial_{xx}f(x_0,y_0)$	$\partial_{xy}f(x_0,y_0)$	$\partial_{yy}f(x_0,y_0)$	$\text{dét}(H_f(x_0,y_0))$	
$(-4,-2)$	$-6e^{-2}$	$8e^{-2}$	$-12e^{-2}$	$8e^{-4}$	maximum
$(0,0)$	2	0	-4	-8	point-selle

24. $f(x,y)=\frac{8}{x}+\frac{x}{y}+y$
    - ⋆ f est de classe $\mathscr{C}^2$ dans son domaine de définition, l'ouvert $\mathbb{R}^2\setminus\{(x,y)\,|\,xy=0\}$.
    - ⋆ 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 :

$$H_f(x,y)=\begin{pmatrix}\frac{16}{x^3}&-\frac{1}{y^2}\\-\frac{1}{y^2}&\frac{2x}{y^3}\end{pmatrix},\qquad\det(H_f(x,y))=\frac{1}{y^3}\left(\frac{16}{x^2}-\frac{1}{y}\right).$$

      $\det(H_f(4,2))>0$ 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 $\mathscr{C}^2$ dans son domaine de définition, l'ouvert $\mathbb{R}^2$.
    - ⋆ Recherche de points critiques :

$$\nabla f(x,y)=\begin{pmatrix}0\\0\end{pmatrix}\iff\begin{cases}2x=0\\\sin(y)=0\end{cases}\iff\iff(x,y)\in\{(0,\kappa\pi)\,|\,\kappa\in\mathbb{Z}\}.$$

      On a une infinité de points critiques qui s'écrivent $(0,\kappa\pi)$ avec $\kappa\in\mathbb{Z}$.
    - ⋆ Nature du point critique :

$$H_f(x,y)=\begin{pmatrix}2&0\\0&\cos(y)\end{pmatrix},\qquad\det(H_f(x,y))=2\cos(y).$$

      $\det(H_f(0,\kappa\pi))=(-1)^\kappa$ et $\partial_{xx}f(0,\kappa\pi)>0$ pour tout$\kappa\in\mathbb{Z}$ donc $(0,\kappa\pi)$ est un minimum si κ est pair et un point-selle si κ est impair.

26. $f(x,y)=(x^2+y^2)e^{-(x^2+y^2)}$.

    On peut remarquer que si on passe aux coordonnées polaire on obtient $w(r)\equiv f(r\cos(\vartheta),r\sin(\vartheta))=r^2e^{-r^2}$, 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 polaire. Si on n'a pas remarqué cette symétrie, on étudie la fonction comme dans les cas précédents :
    - ⋆ f est de classe $\mathscr{C}^2$ 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 secondes 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 $\mathscr{C}^2$ 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 :

$$H_f(x,y)=\begin{pmatrix}6(x-2)&0\\0&6(y+2)\end{pmatrix},\qquad\det(H_f(x,y))=36(x-2)(y+2).$$

dét$(H_f(0,0)) < 0$ donc $(0,0)$ est un point-selle ;
dét$(H_f(0,-4)) > 0$ et $\partial_{xx} f(0,-4) < 0$ donc $(0,-4)$ est un maximum ;
dét$(H_f(4,0)) > 0$ et $\partial_{xx} f(4,0) > 0$ donc $(4,0)$ est un minimum ;
dét$(H_f(4,-4)) < 0$ $(4,-4)$ est un point-selle.

28. $f(x,y) = (x^2+y^2-y^3)e^{-y}$
   - ⋆ f est de classe $\mathscr{C}^2$ dans son domaine de définition, l'ouvert $\mathbb{R}^2$.
   - ⋆ Recherche de points critiques :

$$\nabla f(x,y) = \begin{pmatrix}0\\0\end{pmatrix} \iff \begin{cases} 2xe^{-y} = 0 \\ (-x^2+2y-4y^2+y^3)e^{-y} = 0 \end{cases} \iff (x,y) \in \left\{ (0,0), (0,2-\sqrt{2}), (0,2+\sqrt{2}) \right\}.$$

   On a quatre trois critiques : $(0,0)$, $(0,2-\sqrt{2})$ et $(0,2+\sqrt{2})$.
   - ⋆ Nature des points critiques :

$$H_f(x,y) = \begin{pmatrix} 2e^{-y} & -2xe^{-y} \\ -2xe^{-y} & (2+x^2-10y+7y^2-y^3)e^{-y} \end{pmatrix}, \qquad \text{dét}(H_f(x,y)) = \left(4-2x^2-20y+14y^2-2y^3\right)e^{-2y}.$$

   dét$(H_f(0,0)) > 0$ et $\partial_{xx} f(0,0) > 0$ donc $(0,0)$ est un minimum ;
   dét$(H_f(0,2-\sqrt{2})) < 0$ donc $(0,2-\sqrt{2})$ est un point-selle ;
   dét$(H_f(0,2+\sqrt{2})) > 0$ et $\partial_{xx} f(0,2+\sqrt{2}) > 0$ donc $(0,2+\sqrt{2})$ est un minimum.

Exercice 6.14

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 $p(x,y) = p_1 x + p_2 y - C(x,y) = 50x - x^2 + 60y - 2y^2 - 2xy$. 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 dét$(H_f(20,5)) = (-2)(-4) - (-2)^2 = 4 > 0$ 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 6.15

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

$$\begin{aligned} f\colon (\mathbb{R}_+^*)^2 &\to \mathbb{R} \\ x &\mapsto f(x,y) = (y-3)(300+6\sqrt{x}-10y) - x - 3000 \end{aligned}$$

La condition nécessaire s'écrit

$$\begin{cases} \partial_x f(x,y) = \frac{3(y-3)}{\sqrt{x}} - 1 = 0 \\ \partial_y f(x,y) = 330 + 6\sqrt{x} - 20y = 0 \end{cases} \quad \Longrightarrow \quad (x_0,y_0) = (164025, 138).$$

La hessienne en ce point est définie négative :

$$\begin{cases} \partial_{xx} f(x,y) = -\frac{3(y-3)}{2\sqrt{x^3}} \\ \partial_{xy} f(x,y) = \frac{3}{\sqrt{x}} \\ \partial_{yy} f(x,y) = \frac{30(y-3)}{\sqrt{x^3}} - \frac{3}{\sqrt{x}} \end{cases} \quad \Longrightarrow \quad \text{dét}(H_f(x_0,y_0)) = -\frac{241}{32805}.$$

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 6.16 (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-contraints) 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

$$\begin{cases} \partial_x f(x,y) = 2\frac{x^4-2}{x^3} = 0 \\ \partial_y f(x,y) = 2\frac{2y^4-1}{y^3} = 0 \end{cases} \quad \Longrightarrow \quad (x_0,y_0) = (\sqrt[4]{2}, 1/\sqrt[4]{2}).$$

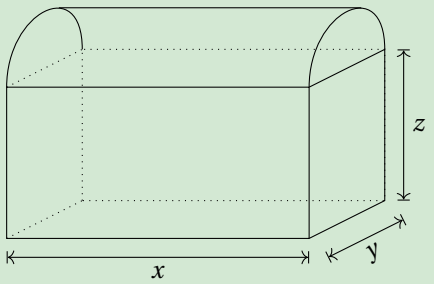
La hessienne en ce point est définie positive :

$$\begin{cases} \partial_{xx} f(x,y) = 2\frac{x^4+6}{x^4} \\ \partial_{xy} f(x,y) = 0 \\ \partial_{yy} f(x,y) = 2\frac{2y^4+3}{y^4} \end{cases} \quad \Longrightarrow \quad \text{dét}(H_f(x_0,y_0)) = 4\frac{2+6}{2}\frac{1+3}{1/2} > 0.$$

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 6.17

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. $V(x,y,z) = xyz + \frac{1}{2}\pi\left(\frac{y}{2}\right)^2 x = xyz + \frac{\pi}{8}xy^2$
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 :

   $$\nabla\tilde{S}(x,y) = \begin{pmatrix} \left(1+\frac{\pi}{4}\right)y - \frac{2C}{x^2} \\ \left(1+\frac{\pi}{4}\right)x - \frac{2C}{y^2} \end{pmatrix} \quad \text{donc} \quad \nabla\tilde{S}(x,y) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff (x,y) = \left(\sqrt[3]{\frac{2C}{1+\frac{\pi}{4}}}, \sqrt[3]{\frac{2C}{1+\frac{\pi}{4}}}\right)$$

   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 :

   $$H_{\tilde{S}}(x,y) = \begin{pmatrix} \frac{4C}{x^3} & 1+\frac{\pi}{4} \\ 1+\frac{\pi}{4} & \frac{4C}{y^3} \end{pmatrix} \quad \text{et} \quad \det(H_{\tilde{S}}(x,y)) = \frac{16C^2}{x^3y^3} - \left(1+\frac{\pi}{4}\right)^2$$

   donc

   $$H_{\tilde{S}}\left(\sqrt[3]{\frac{2C}{1+\frac{\pi}{4}}}, \sqrt[3]{\frac{2C}{1+\frac{\pi}{4}}}\right) = \begin{pmatrix} 2\left(1+\frac{\pi}{4}\right) & 1+\frac{\pi}{4} \\ 1+\frac{\pi}{4} & 2\left(1+\frac{\pi}{4}\right) \end{pmatrix} \quad \text{et} \quad \det\left(H_{\tilde{S}}\left(\sqrt[3]{\frac{2C}{1+\frac{\pi}{4}}}, \sqrt[3]{\frac{2C}{1+\frac{\pi}{4}}}\right)\right) = 3\left(1+\frac{\pi}{4}\right)^2.$$

   On conclut que l'unique point critique est bien un minimum et l'on a $z\left(\sqrt[3]{\frac{2C}{1+\frac{\pi}{4}}}, \sqrt[3]{\frac{2C}{1+\frac{\pi}{4}}}\right) = \frac{C}{\left(\frac{2C}{1+\frac{\pi}{4}}\right)^{2/3}} - \frac{\pi}{8}\left(\frac{2C}{1+\frac{\pi}{4}}\right)^{2/3}$

CHAPITRE 7

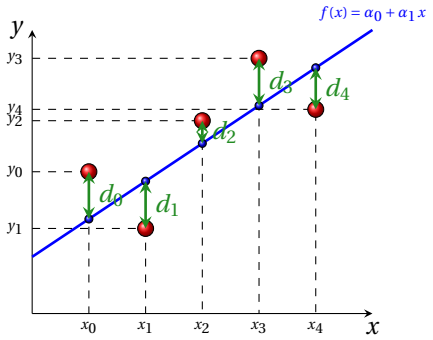
Approximation au sens des moindres carrées : fonction de meilleur approximation (*fitting*)

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 nœuds 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 écarts soit minimale.

Dans tout le chapitre nous considérons un nuage de $n+1$ points $\{(x_i, y_i)\}_{i=0}^{n}$.

7.1. *Fitting* par une relation affine

Supposons que deux grandeurs x et y sont liées approximativement par une relation affine, *i.e.* $f(x) = \alpha_0 + \alpha_1 x$ (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 α_0 et α_1 pour que la droite d'équation $y = \alpha_0 + \alpha_1 x$ s'ajuste *le mieux possible* aux points observés. Pour cela, introduisons $d_i(\alpha_0, \alpha_1) \equiv y_i - (\alpha_0 + \alpha_1 x)$ l'écart vertical du point (x_i, y_i) par rapport à la droite :



La méthode des moindres carrés est celle qui choisit α_0 et α_1 de sorte que *la somme des carrés de ces écarts soit minimale.* Pour cela, on doit minimiser la fonction $\mathscr{E} : \mathbb{R}^2 \to \mathbb{R}_+$ définie par

$$\mathscr{E}(\alpha_0, \alpha_1) = \sum_{i=0}^{n} d_i^2 = \sum_{i=0}^{n} (y_i - \alpha_0 - \alpha_1 x_i)^2.$$

Pour minimiser $\mathscr{E}$ on cherche d'abord les points stationnaires, *i.e.* les points (α_0, α_1) qui vérifient $\frac{\partial \mathscr{E}}{\partial \alpha_0} = \frac{\partial \mathscr{E}}{\partial \alpha_1} = 0$. Puisque

$$\frac{\partial \mathscr{E}}{\partial \alpha_0}(\alpha_0, \alpha_1) = -2\left(\sum_{i=0}^{n} (y_i - \alpha_0 - \alpha_1 x_i)\right), \qquad \frac{\partial \mathscr{E}}{\partial \alpha_1}(\alpha_0, \alpha_1) = -2\left(\sum_{i=0}^{n} x_i (y_i - \alpha_0 - \alpha_1 x_i)\right),$$

alors

$$\begin{cases} \frac{\partial \mathscr{E}}{\partial \alpha_0}(\alpha_0,\alpha_1)=0 \\ \frac{\partial \mathscr{E}}{\partial \alpha_1}(\alpha_0,\alpha_1)=0 \end{cases} \iff \begin{cases} \sum_{i=0}^n (y_i-\alpha_0-\alpha_1 x_i)=0 \\ \sum_{i=0}^n x_i(y_i-\alpha_0-\alpha_1 x_i)=0 \end{cases} \iff \begin{cases} \sum_{i=0}^n y_i-\alpha_0\sum_{i=0}^n 1-\alpha_1\sum_{i=0}^n x_i=0 \\ \sum_{i=0}^n x_i y_i-\alpha_0\sum_{i=0}^n x_i-\alpha_1\sum_{i=0}^n x_i^2=0 \end{cases}$$

$$\iff \begin{cases} (n+1)\alpha_0+\left(\sum_{i=0}^n x_i\right)\alpha_1=\sum_{i=0}^n y_i \\ \left(\sum_{i=0}^n x_i\right)\alpha_0+\left(\sum_{i=0}^n x_i^2\right)\alpha_1=\sum_{i=0}^n y_i x_i \end{cases} \iff \underbrace{\begin{bmatrix} (n+1) & \sum_{i=0}^n x_i \\ \sum_{i=0}^n x_i & \sum_{i=0}^n x_i^2 \end{bmatrix}}_{\mathbb{F}} \underbrace{\begin{bmatrix} a_0 \\ a_1 \end{bmatrix}}_{\mathbf{a}} = \underbrace{\begin{bmatrix} \sum_{i=0}^n y_i \\ \sum_{i=0}^n x_i y_i \end{bmatrix}}_{\mathbf{b}}$$

NB : les points sont indicés de 0 à n ainsi $n+1$ est le nombre de points!

On peut résoudre à la main ce système linéaire et on trouve

$$\begin{cases} \alpha_0=\dfrac{\left(\sum_{i=0}^n x_i\right)\left(\sum_{i=0}^n x_i y_i\right)-\left(\sum_{i=0}^n y_i\right)\left(\sum_{i=0}^n x_i^2\right)}{\left(\sum_{i=0}^n x_i\right)^2-(n+1)\left(\sum_{i=0}^n x_i^2\right)}, \\ \alpha_1=\dfrac{\left(\sum_{i=0}^n x_i\right)\left(\sum_{i=0}^n y_i\right)-(n+1)\left(\sum_{i=0}^n x_i y_i\right)}{\left(\sum_{i=0}^n x_i\right)^2-(n+1)\left(\sum_{i=0}^n x_i^2\right)}. \end{cases}$$

On a trouvé un seul point stationnaire. La fonction étant convexe pour tout (a_0,a_1), on peut conclure qu'il s'agit d'un minimum.

La droite d'équation $y=\alpha_1 x+\alpha_0$ ainsi calculée s'appelle *droite de régression de y par rapport à x*.

EXEMPLE

Soit les 5 points $\{(1,1),(2,2),(3,1),(4,2),(5,3)\}$ (donc $n=4$). On cherche la droite de meilleure approximation $y=\alpha_0+\alpha_1 x$. Il s'agit de chercher α_0 et α_1 solution du système linéaire

$$\begin{pmatrix} (n+1) & \sum_{i=0}^n x_i \\ \sum_{i=0}^n x_i & \sum_{i=0}^n x_i^2 \end{pmatrix}\begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix}=\begin{pmatrix} \sum_{i=0}^n y_i \\ \sum_{i=0}^n x_i y_i \end{pmatrix}$$

$$\implies \begin{pmatrix} 4+1 & 1+2+3+4+5 \\ 1+2+3+4+5 & 1^2+2^2+3^2+4^2+5^2 \end{pmatrix}\begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix}=\begin{pmatrix} 1+2+1+2+3 \\ 1\times 1+2\times 2+3\times 1+4\times 2+5\times 3 \end{pmatrix}$$

$$\implies \begin{pmatrix} 5 & 15 \\ 15 & 55 \end{pmatrix}\begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix}=\begin{pmatrix} 9 \\ 31 \end{pmatrix}$$

Donc $\alpha_0=\frac{3}{5}=0.6$ et $\alpha_1=\frac{2}{5}=0.4$.

```
xp=[1:5];
yp=[1,2,1,2,3];
n=numel(xp)
A=[n, sum(xp); sum(xp), sum(xp.^2) ]
b=[sum(yp); sum(xp.*yp)]
alpha=A\b
f=@(t)[alpha(1)+alpha(2)*t];
xx=linspace(0,2*pi,100);
plot(xp,yp,'o',xx,f(xx))
erreur=sum((yp-f(xp)).^2)
```

Notons que l'élément $(\mathbb{F})_{kj}=\sum_{i=0}^n x_i^{k+j}$ est le produit scalaire du vecteur $(x_0^k,\dots,x_n^k)$ avec le vecteur $(x_0^j,\dots,x_n^j)$ et que l'élément $b_k=\sum_{i=0}^n x_i^k y_i$ est le produit scalaire du vecteur $(x_0^k,\dots,x_n^k)$ avec le vecteur colonne $\mathbf{y}=(y_0,\dots,y_n)$; 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 $k=0,1$:

$$\mathbb{A}\overset{\text{def}}{=}\underbrace{\begin{pmatrix} 1 & x_0 \\ 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}}_{(n+1)\times(m+1)}.$$

En effet,

$$\mathbb{A}^T\mathbb{A} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_0 & x_1 & \dots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_0 \\ 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} n+1 & \sum_{i=0}^n x_i \\ \sum_{i=0}^n x_i & \sum_{i=0}^n x_i^2 \end{pmatrix} \qquad \mathbb{A}^T\mathbf{y} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_0 & x_1 & \dots & x_n \end{pmatrix} \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^n y_i \\ \sum_{i=0}^n x_i y_i \end{pmatrix}$$

7.1.1. *Fitting* linéaire après transformations

Même si la relation entre deux quantités n'est pas linéaire, il est parfois possible d'appliquer une transformation pour trouver une relation linéaire.

***Fitting* linéaire après transformation d'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 $\ln(f(x))=\ln(a)+kx$ qui est linéaire et a la forme $\alpha_0+\alpha_1 x$ avec $\alpha_1=k$ et $\alpha_0=\ln(a)$.

On peut alors calculer l'équation de la droite de régression sur l'ensemble $\{(x_i,\ln(y_i))\}_{i=0}^n$ et obtenir ainsi k et $\ln(a)$. [1]

***Fitting* linéaire après transformation d'une puissance** Soit $a>0$ et considérons la fonction $f(x)=ax^k$: elle est non-linéaire mais si on prend son logarithme on obtient $\ln(f(x))=\ln(a)+k\ln(x)$ qui est linéaire et a la forme $\alpha_0+\alpha_1 x$ avec $\alpha_1=k$ et $\alpha_0=\ln(a)$.

On peut alors calculer l'équation de la droite de régression sur l'ensemble $\{(\ln(x_i),\ln(y_i))\}_{i=0}^n$ et obtenir ainsi k et $\ln(a)$. [2]

On verra dans la prochaine section comment travailler directement avec la fonction polynomiale f.

Exemple (Fitting linéaire après transformation)
On mesure plusieurs fois la pression P et le volume V d'un gaz de masse donnée. On obtient ainsi $n+1$ mesures $\{P_i,V_i\}_{i=0}^n$. Selon la thermodynamique, ces quantités sont liées par une relation du type $PV^\gamma=C$ où γ et C sont deux constantes à calculer.
On a $\ln(PV^\gamma)=\ln(C)$ ainsi $\ln(P)+\gamma\ln(V)=\ln(C)$. Si on pose $x=\ln(V)$ et $y=\ln(P)$, on a une relation de la forme $y=\alpha_1 x+\alpha_0$ avec $\alpha_1=-\gamma$ et $\alpha_0=\ln(C)$.
On peut alors calculer l'équation de la droite de régression sur l'ensemble $\{(\ln(V_i),\ln(P_i))\}_{i=0}^n$ et obtenir ainsi $\gamma=-\alpha_1$ et $C=e^{\alpha_0}$.

Exemple
Soit les 5 points $\{(1,1),(2,2),(3,1),(4,2),(5,3)\}$ (donc $n=4$). On cherche la fonction de meilleure approximation de la forme $y=Ae^{Bx}$. Si on calcule le logarithme de cette fonction on trouve $\ln(y)=\ln(A)+Bx$. On peut alors calculer la droite de meilleur approximation sur l'ensemble $\{(1,\ln(1)),(2,\ln(2)),(3,\ln(1)),(4,\ln(2)),(5,\ln(3))\}$ et obtenir ainsi B et $\ln(A)$. Notons $\alpha_0=\ln(A)$ et $\alpha_1=B$, il s'agit de chercher α_0 et α_1 solution du système linéaire

$$\begin{pmatrix} (n+1) & \sum_{i=0}^n x_i \\ \sum_{i=0}^n x_i & \sum_{i=0}^n x_i^2 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^n \ln(y_i) \\ \sum_{i=0}^n x_i\ln(y_i) \end{pmatrix}$$

$$\implies \begin{pmatrix} 4+1 & 1+2+3+4+5 \\ 1+2+3+4+5 & 1^2+2^2+3^2+4^2+5^2 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} = \begin{pmatrix} \ln(1)+\ln(2)+\ln(1)+\ln(2)+\ln(3) \\ 1\times\ln(1)+2\times\ln(2)+3\times\ln(1)+4\times\ln(2)+5\times\ln(3) \end{pmatrix}$$

$$\implies \begin{pmatrix} 5 & 15 \\ 15 & 55 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} = \begin{pmatrix} 2\ln(2)+\ln(3) \\ 6\ln(2)+5\ln(3) \end{pmatrix}$$

$$\implies \begin{pmatrix} 5 & 15 \\ 0 & 10 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} = \begin{pmatrix} 2\ln(2)+\ln(3) \\ 2\ln(3) \end{pmatrix}$$

Donc $\alpha_1=\frac{\ln(3)}{5}$ et $\alpha_1=\frac{\ln(4)-\ln(9)}{5}$ et enfin $B=\alpha_1$ et $A=e^{\alpha_0}$.

```
xp=[1:5];
yp=[1,2,1,2,3];
ylp=log(yp);
xx=linspace(0,6,100);
%sol=[5 15;15 55]\[sum(ylp);sum(xp.*ylp)]
```

1. 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 calcule la droite de régression sur l'ensemble $\{(x_i,\ln(y_i))\}_{i=0}^n$ on minimise D_i et non d_i.
2. À nouveau, 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 $d_i=y_i-ax_i^k$ et $D_i=\ln(y_i)-(k\ln(x_i)+\ln(a))$, lorsqu'on calcule la droite de régression sur l'ensemble $\{(\ln(x_i),\ln(y_i))\}_{i=0}^n$ on minimise D_i et non d_i.

```
%alpha0=sol(1);
%alpha1=sol(2);
alpha0=2*(log(2)-log(3))/5;
alpha1=log(3)/5;
A=exp(alpha0);
B=alpha1;
subplot(1,2,1)
f=@(t)[alpha0+alpha1*t];
plot(xp,ylp,'o',xx,f(xx))
subplot(1,2,2)
f=@(t)[A*exp(B*t)];
plot(xp,yp,'o',xx,f(xx))
```

7.2. 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 $d_i(a_0,a_1,\dots,a_m)=y_i-\left(\sum_{j=0}^m a_j x_i^j\right)$ 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 $\mathscr{E}:\mathbb{R}^{m+1}\to\mathbb{R}_+$ définie par

$$\mathscr{E}(a_0,a_1,a_2,\dots,a_m)=\sum_{i=0}^n\left(y_i-f(x_i)\right)^2=\sum_{i=0}^n\left(y_i-\sum_{j=0}^m a_j x_i^j\right)^2=\sum_{i=0}^n\left(y_i-a_0-a_1x_i-a_2x_i^2\cdots-a_mx_i^m\right)^2.$$

Pour minimiser $\mathscr{E}$ on cherche d'abord ses points stationnaires, *i.e.* les points qui vérifient $\frac{\partial\mathscr{E}}{\partial a_j}=0$ pour $j=0,\dots,m$. Puisque

$$\begin{aligned}
\frac{\partial\mathscr{E}}{\partial a_0}(a_0,a_1,a_2,\dots,a_m)&=-2\sum_{i=0}^n x_i^0\left(y_i-\sum_{j=0}^m a_j x_i^j\right)=-2\sum_{i=0}^n x_i^0\left(y_i-a_0-a_1x_i-a_2x_i^2\cdots-a_mx_i^m\right),\\
\frac{\partial\mathscr{E}}{\partial a_1}(a_0,a_1,a_2,\dots,a_m)&=-2\sum_{i=0}^n x_i^1\left(y_i-\sum_{j=0}^m a_j x_i^j\right)=-2\sum_{i=0}^n x_i^1\left(y_i-a_0-a_1x_i-a_2x_i^2\cdots-a_mx_i^m\right),\\
\frac{\partial\mathscr{E}}{\partial a_2}(a_0,a_1,a_2,\dots,a_m)&=-2\sum_{i=0}^n x_i^2\left(y_i-\sum_{j=0}^m a_j x_i^j\right)=-2\sum_{i=0}^n x_i^2\left(y_i-a_0-a_1x_i-a_2x_i^2\cdots-a_mx_i^m\right),\\
&\vdots\\
\frac{\partial\mathscr{E}}{\partial a_m}(a_0,a_1,a_2,\dots,a_m)&=-2\sum_{i=0}^n\left(x_i^m\left(y_i-\sum_{j=0}^m a_j x_i^j\right)\right)=-2\sum_{i=0}^n x_i^m\left(y_i-a_0-a_1x_i-a_2x_i^2\cdots-a_mx_i^m\right),
\end{aligned}$$

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

$$\begin{cases}
\frac{\partial\mathscr{E}}{\partial a_0}(a_0,a_1,a_2,\dots,a_m)=0\\
\frac{\partial\mathscr{E}}{\partial a_1}(a_0,a_1,a_2,\dots,a_m)=0\\
\frac{\partial\mathscr{E}}{\partial a_2}(a_0,a_1,a_2,\dots,a_m)=0\\
\vdots\\
\frac{\partial\mathscr{E}}{\partial a_m}(a_0,a_1,a_2,\dots,a_m)=0
\end{cases}
\iff
\begin{cases}
a_0(n+1)+a_1\sum_{i=0}^n x_i+a_2\sum_{i=0}^n x_i^2\cdots+a_m\sum_{i=0}^n x_i^m=\sum_{i=0}^n y_i\\
a_0\sum_{i=0}^n x_i+a_1\sum_{i=0}^n x_i^2+a_2\sum_{i=0}^n x_i^3\cdots+a_m\sum_{i=0}^n x_i^{m+1}=\sum_{i=0}^n y_ix_i\\
a_0\sum_{i=0}^n x_i^2+a_1\sum_{i=0}^n x_i^3+a_2\sum_{i=0}^n x_i^4\cdots+a_m\sum_{i=0}^n x_i^{m+2}=\sum_{i=0}^n y_ix_i^2\\
\vdots\\
a_0\sum_{i=0}^n x_i^m+a_1\sum_{i=0}^n x_i^{m+1}+a_2\sum_{i=0}^n x_i^{m+2}\cdots+a_m\sum_{i=0}^n x_i^{2m}=\sum_{i=0}^n y_ix_i^m
\end{cases}$$

$$\iff
\underbrace{\begin{bmatrix}
(n+1) & \sum_{i=0}^n x_i & \sum_{i=0}^n x_i^2 & \cdots & \sum_{i=0}^n x_i^m\\
\sum_{i=0}^n x_i & \sum_{i=0}^n x_i^2 & \sum_{i=0}^n x_i^3 & \cdots & \sum_{i=0}^n x_i^{m+1}\\
\sum_{i=0}^n x_i^2 & \sum_{i=0}^n x_i^3 & \sum_{i=0}^n x_i^4 & \cdots & \sum_{i=0}^n x_i^{m+2}\\
\vdots & \vdots & & \vdots & \\
\sum_{i=0}^n x_i^m & \sum_{i=0}^n x_i^{m+1} & \sum_{i=0}^n x_i^{m+2} & \cdots & \sum_{i=0}^n x_i^{2m}
\end{bmatrix}}_{\mathbb{F}}
\underbrace{\begin{bmatrix}a_0\\a_1\\a_2\\\vdots\\a_m\end{bmatrix}}_{\mathbf{a}}
=
\underbrace{\begin{bmatrix}\sum_{i=0}^n y_i\\\sum_{i=0}^n y_ix_i\\\sum_{i=0}^n y_ix_i^2\\\vdots\\\sum_{i=0}^n y_ix_i^m\end{bmatrix}}_{\mathbf{b}}$$

Quand $m \geq n$, le polynôme de meilleure approximation coïncide avec le polynôme d'interpolation de $\mathbb{R}_n[x]$.

EXEMPLE
Soit les 5 points $\{(1,1),(2,2),(3,1),(4,2),(5,3)\}$ (donc $n=4$). On cherche la parabole de meilleure approximation $y=\alpha_0+\alpha_1 x+\alpha_2 x^2$. Il s'agit de chercher α_0, α_1 et α_2 solution du système linéaire

$$\begin{pmatrix} (n+1) & \sum_{i=0}^n x_i & \sum_{i=0}^n x_i^2 \\ \sum_{i=0}^n x_i & \sum_{i=0}^n x_i^2 & \sum_{i=0}^n x_i^3 \\ \sum_{i=0}^n x_i^2 & \sum_{i=0}^n x_i^3 & \sum_{i=0}^n x_i^4 \end{pmatrix}\begin{pmatrix}\alpha_0\\ \alpha_1\\ \alpha_2\end{pmatrix} = \begin{pmatrix} \sum_{i=0}^n y_i \\ \sum_{i=0}^n x_i y_i \\ \sum_{i=0}^n x_i^2 y_i \end{pmatrix}$$

$$\Longrightarrow \quad \begin{pmatrix} 4+1 & 1+2+3+4+5 & 1^2+2^2+3^2+4^2+5^2 \\ 1+2+3+4+5 & 1^2+2^2+3^2+4^2+5^2 & 1^3+2^3+3^3+4^3+5^3 \\ 1^2+2^2+3^2+4^2+5^2 & 1^3+2^3+3^3+4^3+5^3 & 1^4+2^4+3^4+4^4+5^4 \end{pmatrix}\begin{pmatrix}\alpha_0\\ \alpha_1\\ \alpha_2\end{pmatrix} = \begin{pmatrix} 1+2+1+2+3 \\ 1\times 1+2\times 2+3\times 1+4\times 2+5\times 3 \\ 1^2\times 1+2^2\times 2+3^2\times 1+4^2\times 2+5^2\times 3 \end{pmatrix}$$

$$\Longrightarrow \quad \begin{pmatrix} 5 & 15 & 55 \\ 15 & 55 & 225 \\ 55 & 225 & 979 \end{pmatrix}\begin{pmatrix}\alpha_0\\ \alpha_1\\ \alpha_2\end{pmatrix} = \begin{pmatrix} 9 \\ 31 \\ 125 \end{pmatrix}$$

Donc $\alpha_0=\frac{8}{5}$, $\alpha_1=-\frac{16}{35}$ et $\alpha_2=\frac{1}{7}$.

```
xp=[1:5];
yp=[1,2,1,2,3];
xx=linspace(0,2*pi,100);
f=@(t)[8/5-16/35*t+1/7*t.^2];
plot(xp,yp,'o',xx,f(xx))
```

Notons que l'élément $(\mathbb{F})_{kj}=\sum_{i=0}^n x_i^{k+j}$ est le produit scalaire du vecteur $(x_0^k, x_1^k, \ldots, x_n^k)$ avec le vecteur $(x_0^j, x_1^j, \ldots, x_n^j)$ et que l'élément $b_k=\sum_{i=0}^n x_i^k y_i$ est le produit scalaire du vecteur $(x_0^k, x_1^k, \ldots, x_n^k)$ avec le vecteur $\mathbf{y}=(y_0, y_1, \ldots, y_n)$; 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 $i=0,\ldots,n$ et $k=0,\ldots m$:

$$\mathbb{A} \overset{\text{def}}{=} \underbrace{\begin{pmatrix} 1 & x_0 & \ldots & x_0^m \\ 1 & x_1 & \ldots & x_1^m \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \ldots & x_n^m \end{pmatrix}}_{(n+1)\times(m+1)}.$$

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.

7.3. 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 $\{\phi_j,\ j=0,\ldots,m\}$. 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 $d_i(a_0,a_1,\ldots,a_m)=y_i-\left(\sum_{j=0}^m a_j\phi_j(x)\right)$ 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 écarts soit minimale.
Pour cela, on doit minimiser la fonction $\mathscr{E}:\mathbb{R}^{m+1}\to\mathbb{R}_+$ définie par

$$\mathscr{E}(a_0,a_1,\ldots,a_m)=\sum_{i=0}^n\left(y_i-\sum_{j=0}^m a_j\phi_j(x_i)\right)^2=\sum_{i=0}^n\left(y_i-a_0\phi_0(x_i)-a_1\phi_1(x_i)-a_2\phi_2(x_i)\cdots-a_m\phi_m(x_i)\right)^2.$$

Pour minimiser $\mathscr{E}$ on cherche d'abord ses points stationnaires, *i.e.* les points qui vérifient $\frac{\partial \mathscr{E}}{\partial a_j} = 0$ pour $j = 0, \dots, m$. Puisque

$$\begin{aligned}
\frac{\partial \mathscr{E}}{\partial a_0}(a_0, a_1, \dots, a_m) &= -2\sum_{i=0}^{n}\left(\phi_0(x_i)\left(y_i - \sum_{j=0}^{m} a_j\phi_j(x_i)\right)\right) = -2\sum_{i=0}^{n}\left(\phi_0(x_i)\left(y_i - a_0\phi_0(x_i) - a_1\phi_1(x_i) - a_2\phi_2(x_i)\cdots - a_m\phi_m(x_i)\right)\right),\\
\frac{\partial \mathscr{E}}{\partial a_1}(a_0, a_1, \dots, a_m) &= -2\sum_{i=0}^{n}\left(\phi_1(x_i)\left(y_i - \sum_{j=0}^{m} a_j\phi_j(x_i)\right)\right) = -2\sum_{i=0}^{n}\left(\phi_1(x_i)\left(y_i - a_0\phi_0(x_i) - a_1\phi_1(x_i) - a_2\phi_2(x_i)\cdots - a_m\phi_m(x_i)\right)\right),\\
&\vdots\\
\frac{\partial \mathscr{E}}{\partial a_m}(a_0, a_1, \dots, a_m) &= -2\sum_{i=0}^{n}\left(\phi_m(x_i)\left(y_i - \sum_{j=0}^{m} a_j\phi_j(x_i)\right)\right) - 2\sum_{i=0}^{n}\left(\phi_m(x_i)\left(y_i - a_0\phi_0(x_i) - a_1\phi_1(x_i) - a_2\phi_2(x_i)\cdots - a_m\phi_m(x_i)\right)\right),
\end{aligned}$$

on obtient alors le système linéaire de $(m+1)$ équations en les $(m+1)$ inconnues $a_0, a_1, \dots, a_m$ suivant

$$\begin{cases}
\frac{\partial \mathscr{E}}{\partial a_0}(a_0, a_1, \dots, a_m) = 0\\
\frac{\partial \mathscr{E}}{\partial a_1}(a_0, a_1, \dots, a_m) = 0\\
\vdots\\
\frac{\partial \mathscr{E}}{\partial a_m}(a_0, a_1, \dots, a_m) = 0
\end{cases}$$

$$\iff \begin{cases}
a_0\sum_{i=0}^{n}\phi_0(x_i)\phi_0(x_i) + a_1\sum_{i=0}^{n}\phi_0(x_i)\phi_1(x_i) + \cdots + a_m\sum_{i=0}^{n}\phi_0(x_i)\phi_m(x_i) = \sum_{i=0}^{n}\phi_0(x_i)y_i\\
a_0\sum_{i=0}^{n}\phi_1(x_i)\phi_0(x_i) + a_1\sum_{i=0}^{n}\phi_1(x_i)\phi_1(x_i) + \cdots + a_m\sum_{i=0}^{n}\phi_1(x_i)\phi_m(x_i) = \sum_{i=0}^{n}\phi_1(x_i)y_i\\
\vdots\\
a_0\sum_{i=0}^{n}\phi_m(x_i)\phi_0(x_i) + a_1\sum_{i=0}^{n}\phi_m(x_i)\phi_1(x_i) + \cdots + a_m\sum_{i=0}^{n}\phi_m(x_i)\phi_m(x_i) = \sum_{i=0}^{n}\phi_m(x_i)y_i
\end{cases}$$

$$\iff \begin{pmatrix}
\sum_{i=0}^{n}\phi_0(x_i)\phi_0(x_i) & \sum_{i=0}^{n}\phi_0(x_i)\phi_1(x_i) & \dots & \sum_{i=0}^{n}\phi_0(x_i)\phi_m(x_i)\\
\sum_{i=0}^{n}\phi_1(x_i)\phi_0(x_i) & \sum_{i=0}^{n}\phi_1(x_i)\phi_1(x_i) & \dots & \sum_{i=0}^{n}\phi_1(x_i)\phi_m(x_i)\\
\vdots & \vdots & & \vdots\\
\sum_{i=0}^{n}\phi_m(x_i)\phi_0(x_i) & \sum_{i=0}^{n}\phi_m(x_i)\phi_1(x_i) & \dots & \sum_{i=0}^{n}\phi_m(x_i)\phi_m(x_i)
\end{pmatrix}\begin{pmatrix} a_0\\ a_1\\ \vdots\\ a_m\end{pmatrix} = \begin{pmatrix} \sum_{i=0}^{n}\phi_0(x_i)y_i\\ \sum_{i=0}^{n}\phi_1(x_i)y_i\\ \vdots\\ \sum_{i=0}^{n}\phi_m(x_i)y_i\end{pmatrix}$$

Si on note $\Phi_{kj} \overset{\text{def}}{=} \sum_{i=0}^{n}\phi_k(x_i)\phi_j(x_i)$, 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.

EXEMPLE

Soit les 5 points $\{(0,1), (\pi/2, 2), (\pi, 1), (3\pi/2, 2), (2\pi, 3)\}$ (donc $n = 4$). On cherche la fonction de meilleure approximation dans l'espace vectoriel engendré par $\{\phi_0(x) = 1, \phi_1(x) = \cos(x)\}$, *i.e.* $y = \alpha_0\phi_0(x) + \alpha_1\phi_1(x) = \alpha_0 + \alpha_1\cos(x)$. Il s'agit de chercher α_0 et α_1 solution du système linéaire

$$\begin{aligned}
&\begin{pmatrix}
\sum_{i=0}^{n}\phi_0(x_i)\phi_0(x_i) & \sum_{i=0}^{n}\phi_0(x_i)\phi_1(x_i)\\
\sum_{i=0}^{n}\phi_1(x_i)\phi_0(x_i) & \sum_{i=0}^{n}\phi_1(x_i)\phi_1(x_i)
\end{pmatrix}\begin{pmatrix}\alpha_0\\ \alpha_1\end{pmatrix} = \begin{pmatrix} \sum_{i=0}^{n}\phi_0(x_i)y_i\\ \sum_{i=0}^{n}\phi_1(x_i)y_i\end{pmatrix}\\
\implies &\begin{pmatrix}
\sum_{i=0}^{n}1 & \sum_{i=0}^{n}\cos(x_i)\\
\sum_{i=0}^{n}\cos(x_i) & \sum_{i=0}^{n}\cos^2(x_i)
\end{pmatrix}\begin{pmatrix}\alpha_0\\ \alpha_1\end{pmatrix} = \begin{pmatrix} \sum_{i=0}^{n}y_i\\ \sum_{i=0}^{n}\cos(x_i)y_i\end{pmatrix}\\
\implies &\begin{pmatrix}
4+1 & \cos(0)+\cos(\pi/2)+\cos(\pi)+\cos(3\pi/2)+\cos(2\pi)\\
\cos(0)+\cos(\pi/2)+\cos(\pi)+\cos(3\pi/2)+\cos(2\pi) & \cos^2(0)+\cos^2(\pi/2)+\cos^2(\pi)+\cos^2(3\pi/2)+\cos^2(2\pi)
\end{pmatrix}\begin{pmatrix}\alpha_0\\ \alpha_1\end{pmatrix}\\
&= \begin{pmatrix} 1+2+1+2+3\\ \cos(0)\times 1+\cos(\pi/2)\times 2+\cos(\pi)\times 1+\cos(3\pi/2)\times 2+\cos(2\pi)\times 3\end{pmatrix}\\
\implies &\begin{pmatrix} 5 & 1\\ 1 & 3\end{pmatrix}\begin{pmatrix}\alpha_0\\ \alpha_1\end{pmatrix} = \begin{pmatrix} 9\\ 3\end{pmatrix}
\end{aligned}$$

Donc $\alpha_0 = \frac{12}{7}$ et $\alpha_1 = \frac{3}{7}$.

```
xp=[0,pi/2,pi,3*pi/2,2*pi];
yp=[1,2,1,2,3];
xx=linspace(0,2*pi,100);
f=@(t)[12/7+3/7*cos(t)];
plot(xp,yp,'o',xx,f(xx))
```

Notons que l'élément Φ_{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 :

$$\mathbb{A}\overset{\text{def}}{=}\underbrace{\begin{pmatrix}\phi_0(x_0) & \phi_1(x_0) & \dots & \phi_m(x_0)\\ \phi_0(x_1) & \phi_1(x_1) & \dots & \phi_m(x_1)\\ \vdots & \vdots & & \vdots\\ \phi_0(x_n) & \phi_1(x_n) & \dots & \phi_m(x_n)\end{pmatrix}}_{(n+1)\times(m+1)}.$$

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 $\mathbb{QR}$ ou bien une décomposition en valeurs singulières de la matrice $\mathbb{A}$. Si $n=m$ on trouve un système carrée qui équivaut à la méthode directe d'interpolation.

EXEMPLE

Considérons l'ensemble de 3 points $\{(-2,4),(0,0),(1,1)\}$ (donc $n=2$). On se propose de calculer les fonctions de meilleure approximation avec

1. $f(x)=a_0+a_1x$ ($m=1$ et $\phi_j(x)=x^j$ avec $j=0,1$)
2. $f=a_0+a_1x+a_2x^2$ ($m=2$ et $\phi_j(x)=x^j$ avec $j=0,1,2$)
3. $f(x)=a_0+a_1e^x$ ($m=1$ et $\phi_j(x)=e^{jx}$ avec $j=0,1$)
4. $f(x)=a_0+a_1e^x+a_2e^{2x}$ ($m=2$ 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 $\mathscr{E}(a_0,a_1)=\sum_{i=0}^2(y_i-(a_0+a_1x_i))^2$. Cela impose la résolution du système linéaire

$$\begin{pmatrix}(n+1) & \sum_{i=0}^n x_i\\ \sum_{i=0}^n x_i & \sum_{i=0}^n x_i^2\end{pmatrix}\begin{pmatrix}a_0\\a_1\end{pmatrix}=\begin{pmatrix}\sum_{i=0}^n y_i\\ \sum_{i=0}^n y_ix_i\end{pmatrix}\quad\Longrightarrow\quad\begin{pmatrix}3 & -1\\ -1 & 5\end{pmatrix}\begin{pmatrix}a_0\\a_1\end{pmatrix}=\begin{pmatrix}5\\-7\end{pmatrix}$$

   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 $\mathscr{E}(a_0,a_1,a_2)=\sum_{i=0}^2(y_i-(a_0+a_1x_i+a_2x_i^2))^2$. Cela impose la résolution du système linéaire

$$\begin{pmatrix}(n+1) & \sum_{i=0}^n x_i & \sum_{i=0}^n x_i^2\\ \sum_{i=0}^n x_i & \sum_{i=0}^n x_i^2 & \sum_{i=0}^n x_i^3\\ \sum_{i=0}^n x_i^2 & \sum_{i=0}^n x_i^3 & \sum_{i=0}^n x_i^4\end{pmatrix}\begin{pmatrix}a_0\\a_1\\a_2\end{pmatrix}=\begin{pmatrix}\sum_{i=0}^n y_i\\ \sum_{i=0}^n y_ix_i\\ \sum_{i=0}^n y_ix_i^2\end{pmatrix}\quad i.e.\quad\begin{pmatrix}3 & -1 & 5\\ -1 & 5 & -7\\ 5 & -7 & 17\end{pmatrix}\begin{pmatrix}a_0\\a_1\\a_2\end{pmatrix}=\begin{pmatrix}5\\-7\\17\end{pmatrix}$$

   Donc $a_0=a_1=0$ et $a_2=1$, *i.e.* $f(x)=x^2$. Notons que dans ce cas $\mathscr{E}(0,0,1)=0$: 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 $\mathscr{E}(a_0,a_1)=\sum_{i=0}^2(y_i-a_0-a_1e^{x_i})^2$. Cela impose la résolution du système linéaire

$$\begin{pmatrix}\sum_{i=0}^n\phi_0(x_i)\phi_0(x_i) & \sum_{i=0}^n\phi_0(x_i)\phi_1(x_i)\\ \sum_{i=0}^n\phi_0(x_i)\phi_1(x_i) & \sum_{i=0}^n\phi_1(x_i)\phi_1(x_i)\end{pmatrix}\begin{pmatrix}a_0\\a_1\end{pmatrix}=\begin{pmatrix}\sum_{i=0}^n\phi_0(x_i)y_i\\ \sum_{i=0}^n\phi_1(x_i)y_i\end{pmatrix}\quad\Longrightarrow\quad\begin{pmatrix}\sum_{i=0}^n 1 & \sum_{i=0}^n e^{x_i}\\ \sum_{i=0}^n e^{x_i} & \sum_{i=0}^n e^{2x_i}\end{pmatrix}\begin{pmatrix}a_0\\a_1\end{pmatrix}=\begin{pmatrix}\sum_{i=0}^n y_i\\ \sum_{i=0}^n e^{x_i}y_i\end{pmatrix}$$
$$\Longrightarrow\quad\begin{pmatrix}3 & e^{-2}+1+e\\ e^{-2}+1+e & e^{-4}+1+e^2\end{pmatrix}\begin{pmatrix}a_0\\a_1\end{pmatrix}=\begin{pmatrix}5\\4e^{-2}+0+e\end{pmatrix}$$

   Donc $a_0\approx 2.842$ et $a_1\approx -0.915$.

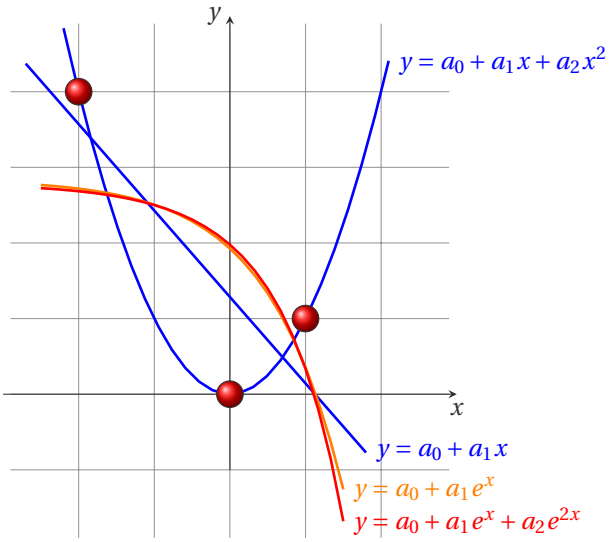
4. Pour $m=2$, il s'agit de chercher a_0, a_1 et a_2 qui minimisent l'erreur $\mathscr{E}(a_0,a_1,a_2)=\sum_{i=0}^2(y_i-a_0-a_1e^{x_i}-a_2e^{2x_i})^2$. Cela impose la résolution du système linéaire

$$\begin{pmatrix} \sum_{i=0}^n \phi_0(x_i)\phi_0(x_i) & \sum_{i=0}^n \phi_0(x_i)\phi_1(x_i) & \sum_{i=0}^n \phi_0(x_i)\phi_2(x_i \\ \sum_{i=0}^n \phi_1(x_i)\phi_0(x_i) & \sum_{i=0}^n \phi_1(x_i)\phi_1(x_i) & \sum_{i=0}^n \phi_1(x_i)\phi_2(x_i) \\ \sum_{i=0}^n \phi_2(x_i)\phi_0(x_i) & \sum_{i=0}^n \phi_2(x_i)\phi_1(x_i) & \sum_{i=0}^n \phi_2(x_i)\phi_2(x_i) \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^n \phi_0(x_i)y_i \\ \sum_{i=0}^n \phi_1(x_i)y_i \\ \sum_{i=0}^n \phi_2(x_i)y_i \end{pmatrix}$$

$$\Longrightarrow \quad \begin{pmatrix} \sum_{i=0}^n 1 & \sum_{i=0}^n e^{x_i} & \sum_{i=0}^n e^{2x_i} \\ \sum_{i=0}^n e^{x_i} & \sum_{i=0}^n e^{2x_i} & \sum_{i=0}^n e^{3x_i} \\ \sum_{i=0}^n e^{x_i} & \sum_{i=0}^n e^{3x_i} & \sum_{i=0}^n e^{4x_i} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^n y_i \\ \sum_{i=0}^n e^{x_i}y_i \\ \sum_{i=0}^n e^{2x_i}y_i \end{pmatrix}$$

$$\Longrightarrow \quad \begin{pmatrix} 3 & e^{-2}+1+e & e^{-4}+1+e^2 \\ e^{-2}+1+e & e^{-4}+1+e^2 & e^{-6}+1+e^3 \\ e^{-4}+1+e^2 & e^{-6}+1+e^3 & e^{-8}+1+e^4 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 5 \\ 4e^{-2}+0+e \\ 4e^{-4}+0+e^2 \end{pmatrix}$$

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 meilleure approximation 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 $\mathscr{E}(a_0,a_1) = \sum_{i=0}^2 (y_i - a_0 - a_1 \frac{1}{x_i})^2$. Cela impose la résolution du système linéaire

$$\begin{pmatrix} \sum_{i=0}^n \phi_0(x_i)\phi_0(x_i) & \sum_{i=0}^n \phi_0(x_i)\phi_1(x_i) \\ \sum_{i=0}^n \phi_0(x_i)\phi_1(x_i) & \sum_{i=0}^n \phi_1(x_i)\phi_1(x_i) \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^n \phi_0(x_i)y_i \\ \sum_{i=0}^n \phi_1(x_i)y_i \end{pmatrix} \quad \Longrightarrow \quad \begin{pmatrix} \sum_{i=0}^n 1 & \sum_{i=0}^n \frac{1}{x_i} \\ \sum_{i=0}^n \frac{1}{x_i} & \sum_{i=0}^n \frac{1}{x_i^2} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^n y_i \\ \sum_{i=0}^n \frac{1}{x_i}y_i \end{pmatrix}$$

$$\Longrightarrow \quad \begin{pmatrix} 3 & \frac{11}{6} \\ \frac{11}{6} & \frac{49}{36} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{5}{3} \end{pmatrix}$$

Donc $a_0 \approx -4.2308$ et $a_1 \approx 6.9231$.

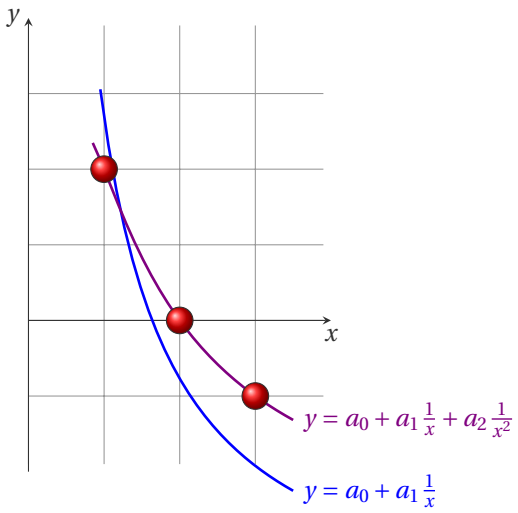
2. Pour $m=2$, il s'agit de chercher a_0, a_1 et a_2 qui minimisent l'erreur $\mathscr{E}(a_0,a_1,a_2) = \sum_{i=0}^2 (y_i - a_0 - a_1 \frac{1}{x_i} - a_2 \frac{1}{x_i^2})^2$. Cela impose la résolution du système linéaire

$$\begin{pmatrix} \sum_{i=0}^n \phi_0(x_i)\phi_0(x_i) & \sum_{i=0}^n \phi_0(x_i)\phi_1(x_i) & \sum_{i=0}^n \phi_0(x_i)\phi_2(x_i \\ \sum_{i=0}^n \phi_1(x_i)\phi_0(x_i) & \sum_{i=0}^n \phi_1(x_i)\phi_1(x_i) & \sum_{i=0}^n \phi_1(x_i)\phi_2(x_i) \\ \sum_{i=0}^n \phi_2(x_i)\phi_0(x_i) & \sum_{i=0}^n \phi_2(x_i)\phi_1(x_i) & \sum_{i=0}^n \phi_2(x_i)\phi_2(x_i) \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^n \phi_0(x_i)y_i \\ \sum_{i=0}^n \phi_1(x_i)y_i \\ \sum_{i=0}^n \phi_2(x_i)y_i \end{pmatrix}$$

$$\Longrightarrow \quad \begin{pmatrix} \sum_{i=0}^n 1 & \sum_{i=0}^n \frac{1}{x_i} & \sum_{i=0}^n \frac{1}{x_i^2} \\ \sum_{i=0}^n \frac{1}{x_i} & \sum_{i=0}^n \frac{1}{x_i^2} & \sum_{i=0}^n \frac{1}{x_i^3} \\ \sum_{i=0}^n \frac{1}{x_i^2} & \sum_{i=0}^n \frac{1}{x_i^3} & \sum_{i=0}^n \frac{1}{x_i^4} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^n y_i \\ \sum_{i=0}^n \frac{1}{x_i^2}y_i \\ \sum_{i=0}^n \frac{1}{x_i^3}y_i \end{pmatrix}$$

$$\Longrightarrow \quad \begin{pmatrix} 3 & \frac{11}{6} & \frac{49}{36} \\ \frac{11}{6} & \frac{49}{36} & \frac{251}{216} \\ \frac{49}{36} & \frac{251}{216} & \frac{1393}{1296} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 1 \\ \frac{5}{3} \\ \frac{17}{9} \end{pmatrix}$$

Donc $a_0 = -\frac{7}{2}$, $a_1 = \frac{17}{2}$ et $a_2 = -3$.



7.4. Résumé

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 générale 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* :

$$\text{minimiser} \qquad \mathscr{E}_f = \sum_{i=0}^{n} \big(y_i - f(x_i)\big)^2.$$

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

$$\mathscr{E}(a_0, a_1) = \sum_{i=0}^{n} \big(y_i - a_0 - a_1 x_i\big)^2$$
$$\nabla\mathscr{E}(a_0, a_1) = \begin{pmatrix} \partial_{a_0}\mathscr{E} \\ \partial_{a_1}\mathscr{E} \end{pmatrix} = \begin{pmatrix} -2\sum_{i=0}^{n} x_i^0 \big(y_i - a_0 - a_1 x_i\big) \\ -2\sum_{i=0}^{n} x_i^1 \big(y_i - a_0 - a_1 x_i\big) \end{pmatrix}$$

  a_0 et a_1 sont alors solution du système linéaire

$$\begin{bmatrix} (n+1) & \sum_{i=0}^{n} x_i \\ \sum_{i=0}^{n} x_i & \sum_{i=0}^{n} x_i^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^{n} y_i \\ \sum_{i=0}^{n} x_i y_i \end{bmatrix}$$

- f polynomiale de degré m, *i.e.* $f(x) = a_0 + a_1 x + \cdots + a_m x^m = \sum_{j=0}^{m} a_j x^j$, ainsi l'erreur est une fonction de $m+1$ variables et l'on a

$$\mathscr{E}(a_0, a_1, a_2, \ldots, a_m) = \sum_{i=0}^{n} \big(y_i - a_0 - a_1 x_i - a_2 x_i^2 \cdots - a_m x_i^m\big)^2 = \sum_{i=0}^{n} \left(y_i - \sum_{j=0}^{m} a_j x_i^j\right)^2$$
$$\nabla\mathscr{E}(a_0, a_1, \ldots, a_m) = \begin{pmatrix} \partial_{a_0}\mathscr{E} \\ \partial_{a_1}\mathscr{E} \\ \partial_{a_2}\mathscr{E} \\ \vdots \\ \partial_{a_m}\mathscr{E} \end{pmatrix} = \begin{pmatrix} -2\sum_{i=0}^{n} x_i^0 \big(y_i - a_0 - a_1 x_i - a_2 x_i^2 \cdots - a_m x_i^m\big) \\ -2\sum_{i=0}^{n} x_i^1 \big(y_i - a_0 - a_1 x_i - a_2 x_i^2 \cdots - a_m x_i^m\big) \\ -2\sum_{i=0}^{n} x_i^2 \big(y_i - a_0 - a_1 x_i - a_2 x_i^2 \cdots - a_m x_i^m\big) \\ \vdots \\ -2\sum_{i=0}^{n} x_i^m \big(y_i - a_0 - a_1 x_i - a_2 x_i^2 \cdots - a_m x_i^m\big) \end{pmatrix}$$

$a_0, a_1, \dots, a_m$ sont alors solution du système linéaire

$$\begin{bmatrix} (n+1) & \sum_{i=0}^n x_i & \sum_{i=0}^n x_i^2 & \dots & \sum_{i=0}^n x_i^m \\ \sum_{i=0}^n x_i & \sum_{i=0}^n x_i^2 & \sum_{i=0}^n x_i^3 & \dots & \sum_{i=0}^n x_i^{m+1} \\ \sum_{i=0}^n x_i^2 & \sum_{i=0}^n x_i^3 & \sum_{i=0}^4 x_i^4 & \dots & \sum_{i=0}^n x_i^{m+2} \\ \vdots & \vdots & & \vdots & \\ \sum_{i=0}^n x_i^m & \sum_{i=0}^n x_i^{m+1} & \sum_{i=0}^n x_i^{m+2} & \dots & \sum_{i=0}^n x_i^{2m} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^n y_i \\ \sum_{i=0}^n y_i x_i \\ \sum_{i=0}^n y_i x_i^2 \\ \vdots \\ \sum_{i=0}^n y_i x_i^m \end{bmatrix}$$

⋆ 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

$$\mathscr{E}(a_0, a_1, a_2, \dots, a_m) = \sum_{i=0}^n \big(y_i - a_0\phi_0(x_i) - a_1\phi_1(x_i) - a_2\phi_2(x_i) \cdots - a_m\phi_m(x_i)\big)^2 = \sum_{i=0}^n \left(y_i - \sum_{j=0}^m a_j\phi_j(x_i)\right)^2$$

$$\nabla\mathscr{E}(a_0, a_1, \dots, a_m) = \begin{pmatrix} \partial_{a_0}\mathscr{E} \\ \partial_{a_1}\mathscr{E} \\ \partial_{a_2}\mathscr{E} \\ \vdots \\ \partial_{a_m}\mathscr{E} \end{pmatrix} = \begin{pmatrix} -2\sum_{i=0}^n \phi_0(x_i)\big(y_i - a_0\phi_0(x_i) - a_1\phi_1(x_i) - a_2\phi_2(x_i) \cdots - a_m\phi_m(x_i)\big) \\ -2\sum_{i=0}^n \phi_1(x_i)\big(y_i - a_0\phi_0(x_i) - a_1\phi_1(x_i) - a_2\phi_2(x_i) \cdots - a_m\phi_m(x_i)\big) \\ -2\sum_{i=0}^n \phi_2(x_i)\big(y_i - a_0\phi_0(x_i) - a_1\phi_1(x_i) - a_2\phi_2(x_i) \cdots - a_m\phi_m(x_i)\big) \\ \vdots \\ -2\sum_{i=0}^n \phi_m(x_i)\big(y_i - a_0\phi_0(x_i) - a_1\phi_1(x_i) - a_2\phi_2(x_i) \cdots - a_m\phi_m(x_i)\big) \end{pmatrix}$$

$a_0, a_1, \dots, a_m$ sont alors solution du système linéaire

$$\begin{pmatrix} \sum_{i=0}^n \phi_0(x_i)\phi_0(x_i) & \sum_{i=0}^n \phi_0(x_i)\phi_1(x_i) & \dots & \sum_{i=0}^n \phi_0(x_i)\phi_m(x_i) \\ \sum_{i=0}^n \phi_1(x_i)\phi_0(x_i) & \sum_{i=0}^n \phi_1(x_i)\phi_1(x_i) & \dots & \sum_{i=0}^n \phi_1(x_i)\phi_m(x_i) \\ \vdots & \vdots & & \vdots \\ \sum_{i=0}^n \phi_m(x_i)\phi_0(x_i) & \sum_{i=0}^n \phi_m(x_i)\phi_1(x_i) & \dots & \sum_{i=0}^n \phi_m(x_i)\phi_m(x_i) \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^n \phi_0(x_i)y_i \\ \sum_{i=0}^n \phi_1(x_i)y_i \\ \vdots \\ \sum_{i=0}^n \phi_m(x_i)y_i \end{pmatrix}$$

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 $\phi_j(x) = e^{jx}$, ou $\phi_j(x) = \cos(jx)$, $\phi_j(x) = x^{-j}\dots$

7.5. Exercices

Exercice 7.1 (Fitting trigonométrique)
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 $\mathscr{E}$ définie par

$$\mathscr{E}\colon \mathbb{R}^2 \to \mathbb{R}_+$$
$$(a, b) \mapsto \mathscr{E}(a, b) = \sum_{i=0}^n d_i^2.$$

Écrire et résoudre le système linéaire qui permet de calculer a et b.

Correction
Pour minimiser $\mathscr{E}$ on cherche ses points stationnaires. Puisque

$$\mathscr{E}(a, b) = \sum_{i=0}^n \left(y_i - \left(a\sin(\frac{\pi}{2}x_i) + b\cos(\frac{\pi}{2}x_i)\right)\right)^2$$

calculons tout d'abord les deux dérivées partielles

$$\frac{\partial\mathscr{E}}{\partial a}(a, b) = -2\left(\sum_{i=0}^n \big(y_i - \big(a\sin\big(\tfrac{\pi}{2}x_i\big) + b\cos\big(\tfrac{\pi}{2}x_i\big)\big)\big)\sin\big(\tfrac{\pi}{2}x_i\big)\right),$$

$$\frac{\partial \mathscr{E}}{\partial b}(a,b) = -2\left(\sum_{i=0}^{n}\left(y_i - \left(a\sin\left(\tfrac{\pi}{2}x_i\right) + b\cos\left(\tfrac{\pi}{2}x_i\right)\right)\right)\cos\left(\tfrac{\pi}{2}x_i\right)\right),$$

et cherchons quand elles s'annulent en même temps. On obtient

$$\begin{cases} \frac{\partial \mathscr{E}}{\partial a}(a,b) = 0 \\ \frac{\partial \mathscr{E}}{\partial b}(a,b) = 0 \end{cases} \iff \begin{cases} \sum_{i=0}^{n}\left(y_i - \left(a\sin\left(\frac{\pi}{2}x_i\right) + b\cos\left(\frac{\pi}{2}x_i\right)\right)\right)\sin(\frac{\pi}{2}x_i) = 0 \\ \sum_{i=0}^{n}\left(y_i - \left(a\sin\left(\frac{\pi}{2}x_i\right) + b\cos\left(\frac{\pi}{2}x_i\right)\right)\right)\cos(\frac{\pi}{2}x_i) = 0 \end{cases}$$

$$\iff \begin{cases} \sum_{i=0}^{n}\left(\left(a\sin\left(\frac{\pi}{2}x_i\right) + b\cos\left(\frac{\pi}{2}x_i\right)\right)\right)\sin(\frac{\pi}{2}x_i) = \sum_{i=0}^{n} y_i \sin(\frac{\pi}{2}x_i) \\ \sum_{i=0}^{n}\left(\left(a\sin\left(\frac{\pi}{2}x_i\right) + b\cos\left(\frac{\pi}{2}x_i\right)\right)\right)\cos(\frac{\pi}{2}x_i) = \sum_{i=0}^{n} y_i \cos(\frac{\pi}{2}x_i) \end{cases}$$

$$\iff \begin{bmatrix} \sum_{i=0}^{n}\sin^2\left(\frac{\pi}{2}x_i\right) & \sum_{i=0}^{n}\sin\left(\frac{\pi}{2}x_i\right)\cos\left(\frac{\pi}{2}x_i\right) \\ \sum_{i=0}^{n}\sin\left(\frac{\pi}{2}x_i\right)\cos\left(\frac{\pi}{2}x_i\right) & \sum_{i=0}^{n}\cos^2\left(\frac{\pi}{2}x_i\right) \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^{n} y_i \sin\left(\frac{\pi}{2}x_i\right) \\ \sum_{i=0}^{n} y_i \cos\left(\frac{\pi}{2}x_i\right) \end{bmatrix}.$$

Si on note

$$U \equiv \sum_{i=0}^{n}\sin^2\left(\tfrac{\pi}{2}x_i\right), \quad V \equiv \sum_{i=0}^{n}\sin\left(\tfrac{\pi}{2}x_i\right)\cos\left(\tfrac{\pi}{2}x_i\right), \quad W \equiv \sum_{i=0}^{n}\cos^2\left(\tfrac{\pi}{2}x_i\right), \quad P \equiv \sum_{i=0}^{n} y_i\sin\left(\tfrac{\pi}{2}x_i\right), \quad Q \equiv \sum_{i=0}^{n} y_i\cos\left(\tfrac{\pi}{2}x_i\right),$$

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

$$a = \frac{WP - VQ}{UW - V^2}, \qquad b = \frac{UQ - VP}{UW - V^2}.$$

Exercice 7.2 (Fitting linéaire avec deux variables)

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 $\mathscr{E}$ définie par

$$\mathscr{E}: \mathbb{R}^3 \to \mathbb{R}_+$$
$$(a,b,c) \mapsto \mathscr{E}(a,b,c) = \sum_{i=0}^{n} d_i^2.$$

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 $\mathscr{E}$ on cherche ses points stationnaires. Puisque

$$\frac{\partial \mathscr{E}}{\partial a}(a,b,c) = -2\left(\sum_{i=0}^{n}(z_i - (a + bx_i + cy_i))\right),$$
$$\frac{\partial \mathscr{E}}{\partial b}(a,b,c) = -2\left(\sum_{i=0}^{n}(z_i - (a + bx_i + cy_i))x_i\right),$$

$$\frac{\partial \mathscr{E}}{\partial c}(a,b,c) = -2\left(\sum_{i=0}^{n}(z_i-(a+bx_i+cy_i))y_i\right),$$

on obtient

$$\begin{cases} \frac{\partial \mathscr{E}}{\partial a}(a,b,c)=0 \\ \frac{\partial \mathscr{E}}{\partial b}(a,b,c)=0 \\ \frac{\partial \mathscr{E}}{\partial c}(a,b,c)=0 \end{cases} \iff \begin{cases} \sum_{i=0}^{n}(z_i-(a+bx_i+cy_i))=0 \\ \sum_{i=0}^{n}(z_i-(a+bx_i+cy_i))x_i=0 \\ \sum_{i=0}^{n}(z_i-(a+bx_i+cy_i))y_i=0 \end{cases} \iff \begin{cases} \sum_{i=0}^{n}(a+bx_i+cy_i)=\sum_{i=0}^{n}z_i \\ \sum_{i=0}^{n}(ax_i+bx_i^2+cy_ix_i)=\sum_{i=0}^{n}z_ix_i \\ \sum_{i=0}^{n}(ay_i+bx_iy_i+cy_i^2)=\sum_{i=0}^{n}z_iy_i \end{cases}$$

$$\iff \begin{pmatrix} (n+1) & \sum_{i=0}^{n}x_i & \sum_{i=0}^{n}y_i \\ \sum_{i=0}^{n}x_i & \sum_{i=0}^{n}x_i^2 & \sum_{i=0}^{n}x_iy_i \\ \sum_{i=0}^{n}y_i & \sum_{i=0}^{n}x_iy_i & \sum_{i=0}^{n}y_i^2 \end{pmatrix}\begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^{n}z_i \\ \sum_{i=0}^{n}z_ix_i \\ \sum_{i=0}^{n}z_iy_i \end{pmatrix}.$$

2. Dans notre cas,

$$\sum_{i=0}^{n}x_i = 7 \qquad \sum_{i=0}^{n}y_i = 4 \qquad \sum_{i=0}^{n}z_i = \frac{11}{2}$$
$$\sum_{i=0}^{n}x_iy_i = 6 \qquad \sum_{i=0}^{n}x_iz_i = \frac{7}{2} \qquad \sum_{i=0}^{n}y_iz_i = \frac{9}{2}$$
$$n+1 = 6 \qquad \sum_{i=0}^{n}x_i^2 = 13 \qquad \sum_{i=0}^{n}y_i^2 = 6$$

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

$$\left(\begin{array}{ccc|c} 6 & 7 & 4 & {}^{11}/_{2} \\ 7 & 13 & 6 & {}^{7}/_{2} \\ 4 & 6 & 6 & {}^{9}/_{2} \end{array}\right) \xrightarrow[L_3\leftarrow L_3-\frac{2}{3}L_1]{L_2\leftarrow L_2-\frac{7}{6}L_1} \left(\begin{array}{ccc|c} 6 & 7 & 4 & {}^{11}/_{2} \\ 0 & {}^{29}/_{6} & {}^{4}/_{3} & -{}^{35}/_{12} \\ 0 & {}^{4}/_{3} & {}^{10}/_{3} & {}^{5}/_{6} \end{array}\right) \xrightarrow{L_3\leftarrow L_3-\frac{8}{29}L_2} \left(\begin{array}{ccc|c} 6 & 7 & 4 & {}^{11}/_{2} \\ 0 & {}^{29}/_{6} & {}^{4}/_{3} & -{}^{35}/_{12} \\ 0 & 0 & {}^{86}/_{29} & {}^{95}/_{58} \end{array}\right)$$

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 7.3 (Fitting polynomial)

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)
  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)
   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
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 7.4 (Fitting non polynomial)

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 $\mathscr{E}: \mathbb{R}^{m+1} \to \mathbb{R}_+$ définie par

$$\mathscr{E}(a_0, a_1, \ldots, a_m) = \sum_{i=0}^{n} \left(y_i - \sum_{j=0}^{m} a_j \phi_j(x_i) \right)^2 .$$

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 $a_0, a_1, \ldots, a_m$ avec $\mathbf{y} = (y_0, y_1, \ldots, y_n)$ et

$$\mathbb{A} \overset{\text{def}}{=} \underbrace{\begin{pmatrix} \phi_0(x_0) & \phi_1(x_0) & \ldots & \phi_m(x_0) \\ \phi_0(x_1) & \phi_1(x_1) & \ldots & \phi_m(x_1) \\ \vdots & \vdots & & \vdots \\ \phi_0(x_n) & \phi_1(x_n) & \ldots & \phi_m(x_n) \end{pmatrix}}_{n \times m}.$$

Choisissons les fonctions ϕ_j de la forme $\phi_j(x) = \sin(j\pi x)$, $j = 0, \ldots, m$. 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|')
```

☆ Exercice 7.5 (Interpolation et *fitting*, utilisation de `function` vues en TP)
Dans cet exercice on fera appel à la `function` `fittingpolynomial.m` écrite en TP pour calculer le fitting polynomial et à l'une des `function` 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 `fittingpolynomiale(P,x,deg)`.

Correction
En TP, dans les fichiers `fittingpolynomialPoly.m` et `naivePoly.m` on a écrit les `function`

```
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,l);
   V(:,2:l) = P(:,1).^(1:l-1);
   alpha = V\P(:,2);
end
```

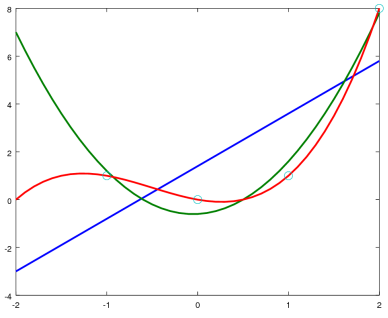
qui calculent les coefficients des polynômes de régression et d'interpolation.
Ensuite, dans les fichiers `fittingpolynomialEval.m` et `naiveEval.m` on a écrit les `function` pour évaluer les polynômes ainsi construit en un vecteur de points :

```
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,

$$p_{\text{linéaire}}(x) = \alpha_0 + \alpha_1 x$$

$$p_{\text{parabolique}}(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$
$$p_{\text{interpolation}}(x) = \frac{x(x-1)(x-2)}{-6} + \frac{(x+1)x(x-2)}{-2} + 8\frac{(x+1)x(x-1)}{6}$$

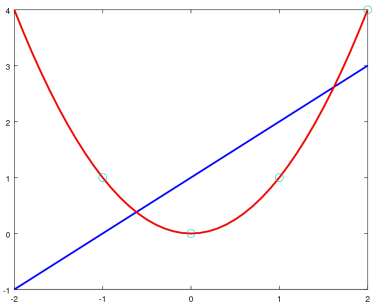
avec

$$\begin{pmatrix} 4 & 2 \\ 2 & 6 \end{pmatrix}\begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} = \begin{pmatrix} 10 \\ 16 \end{pmatrix} \qquad \begin{pmatrix} 4 & 2 & 6 \\ 2 & 6 & 8 \\ 6 & 8 & 18 \end{pmatrix}\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 10 \\ 16 \\ 34 \end{pmatrix}$$

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,

$$p_{\text{linéaire}}(x) = \alpha_0 + \alpha_1 x$$
$$p_{\text{parabolique}}(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$
$$p_{\text{interpolation}}(x) = \frac{x(x-1)(x-2)}{-6} + \frac{(x+1)x(x-2)}{-2} + 4\frac{(x+1)x(x-1)}{6} = x^2$$

avec

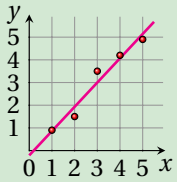
$$\begin{pmatrix} 4 & 2 \\ 2 & 6 \end{pmatrix}\begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} = \begin{pmatrix} 6 \\ 8 \end{pmatrix} \qquad \begin{pmatrix} 4 & 2 & 6 \\ 2 & 6 & 8 \\ 6 & 8 & 18 \end{pmatrix}\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 6 \\ 8 \\ 18 \end{pmatrix}$$

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 7.6 (Fitting linéaire)

Calculer la droite de meilleur 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

Nous avons 5 points, ainsi $n = 4$.

Il s'agit de chercher a_0 et a_1 qui minimisent l'erreur $\mathscr{E}(a_0, a_1) = \sum_{i=0}^{n}(y_i - (a_0 + a_1 x_i))^2$. Cela impose la résolution du système linéaire

$$\begin{pmatrix} (n+1) & \sum_{i=0}^{n} x_i \\ \sum_{i=0}^{n} x_i & \sum_{i=0}^{n} x_i^2 \end{pmatrix}\begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^{n} y_i \\ \sum_{i=0}^{n} y_i x_i \end{pmatrix} \implies \begin{pmatrix} 5 & 15 \\ 15 & 55 \end{pmatrix}\begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 15 \\ 55.7 \end{pmatrix}$$

Donc $a_0 = -0.21$ et $a_1 = 1.07$.
En utilisant les `function` de l'exercice 7.3 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 7.7 (Fitting parabolique)
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
Nous avons 8 points donc $n = 7$.

1. La droite de régression a équation $y = a_0 + a_1 x$ avec a_0, a_1 solution du système linéaire

$$\begin{pmatrix} 8 & \sum_{i=0}^{7} x_i \\ \sum_{i=0}^{7} x_i & \sum_{i=0}^{7} x_i^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^{7} y_i \\ \sum_{i=0}^{7} y_i x_i \end{pmatrix} \quad i.e. \quad \begin{pmatrix} 8 & 19.8 \\ 19.8 & 58.48 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 137.364 \\ 429.4061 \end{pmatrix}$$

   et on obtient

$$\begin{cases} a_0 = -6.189895251, \\ a_1 = 9.438543536. \end{cases}$$

   L'erreur est

$$\sum_{i=0}^{7} (y_i - (a_0 + a_1 x_i))^2 = 30.20147192.$$

2. 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

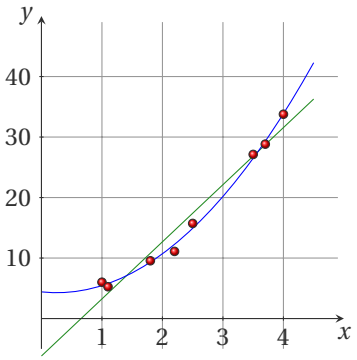
$$\begin{pmatrix} 8 & \sum_{i=0}^{7} x_i & \sum_{i=0}^{7} x_i^2 \\ \sum_{i=0}^{7} x_i & \sum_{i=0}^{7} x_i^2 & \sum_{i=0}^{7} x_i^3 \\ \sum_{i=0}^{7} x_i^2 & \sum_{i=0}^{7} x_i^3 & \sum_{i=0}^{7} x_i^4 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^{7} y_i \\ \sum_{i=0}^{7} y_i x_i \\ \sum_{i=0}^{7} y_i x_i^2 \end{pmatrix} \quad i.e. \quad \begin{pmatrix} 8 & 19.8 & 58.48 \\ 19.8 & 58.48 & 191.964 \\ 58.48 & 191.964 & 668.9284 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 137.364 \\ 429.4061 \\ 1462.63437 \end{pmatrix}$$

   et on obtient

$$\begin{cases} a_0 = 4.40567376946090050, \\ a_1 = -1.06889613092575431, \\ a_2 = 2.10811821540086797. \end{cases}$$

   L'erreur est

$$\sum_{i=0}^{7} (y_i - (a_0 + a_1 x_i + a_2 x_i^2))^2 = 3.304259349.$$



Exercice 7.8 (Fitting parabolique)
Le tableau ci-dessous donne la conductivité thermique k du sodium pour différentes valeurs de la température. On veut calculer la parabole de meilleur 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

$$\begin{pmatrix} 6 & \sum_{i=0}^4 x_i & \sum_{i=0}^4 x_i^2 \\ \sum_{i=0}^4 x_i & \sum_{i=0}^4 x_i^2 & \sum_{i=0}^4 x_i^3 \\ \sum_{i=0}^4 x_i^2 & \sum_{i=0}^4 x_i^3 & \sum_{i=0}^4 x_i^4 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^4 y_i \\ \sum_{i=0}^4 y_i x_i \\ \sum_{i=0}^4 y_i x_i^2 \end{pmatrix} \quad \textit{i.e.} \quad \begin{pmatrix} 8 & 16.1 & 44.79 \\ 16.1 & 44.79 & 141.311 \\ 44.79 & 141.311 & 481.5123 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 108.536 \\ 322.7425 \\ 1067.97905 \end{pmatrix}$$

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 7.9 (Fitting polynomial)
La viscosité cinématique μ 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
μ (10^{-3} m^2 s^{-1})	1.79	1.13	0.696	0.519	0.338	0.321	0.296

On veut évaluer les valeurs $\mu(10°)$, $\mu(30°)$, $\mu(60°)$, $\mu(90°)$ 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

$$r(10°) = 1.004300740 \quad r(30°) = 0.9114735501 \quad r(60°) = 0.9114735501 \quad r(90°) = 0.249145396$$

☆ Exercice 7.10
Considérons $n = 10$ points $P_i = (x_i, y_i)$ avec $x_i = (i-1)/(n-1)$, $i = 1, \ldots, n$ 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 evaluera les polynomes 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 functions")
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 7.11
L'espérance de vie dans un pays a évoluée 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 7.12 (Fitting linéaire après transformation)

L'évolution de la concentration c d'un médicament dans le sang en fonction du temps t est décrite par la fonction $f(t) = Ate^{Bt}$. En utilisant les mesures suivantes et une transformation adéquate de f estimer A et B par régression linéaire :

t	1	2	3	4	5	6	7	8
c	8.0	12.3	15.5	16.8	17.1	15.8	15.2	14.0

Correction

On a $\ln(f(t)) = \ln(A) + \ln(t) + Bt$ ainsi $\ln(f(t)) - \ln(t) = \ln(A) + Bt$ qui est linéaire en B et a la forme $\alpha_0 + \alpha_1 t$ avec $\alpha_1 = B$ et $\alpha_0 = \ln(A)$. On peut alors calculer l'équation de la droite de régression sur l'ensemble $\left\{(t_i, y_i = \ln(c_i) - \ln(t_i))\right\}_{i=0}^{n}$ et obtenir ainsi B et $\ln(A)$.

$$\begin{pmatrix} 8 & \sum_{i=0}^{7} t_i \\ \sum_{i=0}^{7} t_i & \sum_{i=0}^{7} t_i^2 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^{7} \ln(c_i) - \ln(t_i) \\ \sum_{i=0}^{7} t_i \left(\ln(c_i) - \ln(t_i)\right) \end{pmatrix}$$

Donc $\alpha_0 \simeq -16.3929$ et $\alpha_1 \simeq 3.8929$ et enfin $B = \alpha_1$ et $A = e^{\alpha_0}$.

```
tp=[1:8];
cp=[8.0,12.3,15.5,16.8,17.1,15.8,15.2,14.0];
```

```
yp=log(cp)-log(tp);
A=[8,sum(tp);sum(tp),sum(tp.^2)]
b=[sum(yp); sum(tp.*yp)]
a=A\b
xx=linspace(0,9,100);
subplot(1,2,1)
f=@(t)[a(1)+a(2)*t];
plot(tp,yp,'o',xx,f(xx))
subplot(1,2,2)
f=@(t)[exp(a(1)).*t.*exp(a(2)*t)];
plot(tp,cp,'o',xx,f(xx))
```

CHAPITRE 8

Statistique descriptive

La statistique descriptive a pour but de décrire, classer et simplifier des données qui peuvent être volumineuses, de les représenter de manière synthétique sous forme de tableaux ou de graphiques, et d'extraire quelques valeurs importantes qui décrivent les propriétés essentielles des données telles que la moyenne, la variance, la corrélation etc.

8.1. Vocabulaire

- **Population** L'ensemble sur lequel porte l'activité statistique s'appelle la population, généralement notée Ω. Lorsque la population est finie, le nombre d'éléments contenus dans Ω est noté N. Les éléments qui constituent la population sont appelés les individus ou encore les unités statistiques.
- **Échantillon** Un échantillon, noté généralement S (pour "*sample*") est une partie de la population prélevée soit de façon aléatoire soit de façon non aléatoire. Le nombre d'éléments de S est noté n.
- **Caractères** Les caractéristiques étudiées sur les individus d'une population sont appelées les caractères. Soit $\mathscr{C}$ l'ensemble des valeurs possibles du caractère, on définit alors un caractères comme une application $\chi : \Omega \to \mathscr{C}$ qui associe à chaque individu $\omega \in \Omega$ la valeur $\chi(\omega) \in \mathscr{C}$ que prend ce caractère sur l'individu ω.

  Il existe deux types de caractères :
  - ⋆ caractères **quantitatifs** : c'est un caractère dont les issues produisent un nombre (caractères simples ou univariés, $\mathscr{C} \subset \mathbb{R}$) ou une suite de nombres (caractères multiples ou multivariés, $\mathscr{C} \subset \mathbb{R}^m$). Parmi les caractères quantitatifs il faut distinguer
    - ⋆ les caractères quantitatifs **continus** qui peuvent prendre toutes les valeurs d'un intervalle,
    - ⋆ les caractères quantitatifs **discrets** qui ne prennent que des valeurs isolées ;
  - ⋆ caractères **qualitatifs** : c'est un caractère dont les issues ne sont pas quantifiables numériquement. On parle alors de modalités et non d'issues dans ce cas. Parmi les caractères qualitatifs il faut distinguer
    - ⋆ les caractères qualitatifs **ordinaux** qui peuvent être ordonnées,
    - ⋆ les caractères qualitatifs **nominaux**.

EXEMPLE
- ⋆ La masse d'un individu est un caractère quantitatif univarié continu ($\mathscr{C} \subset \mathbb{R}^+$).
- ⋆ Le relevé de températures d'une ville pendant le mois de juin est un caractère quantitatif multivarié continu ($\mathscr{C} \subset \mathbb{R}^{30}$).
- ⋆ Le genre est un caractère qualitatif nominal ($\mathscr{C} = \{\text{homme}, \text{femme}\}$). On peut bien sûr coder la valeur "homme" par "0" et "femme" par "1" mais cela ne donne ni un sens à l'ordre ni le transforme en un caractère quantitatif.

8.1.1. Série statistique et distribution statistique non groupée

Considérons une série statistique associée à un caractère, c'est-à-dire un échantillon de n valeurs réelles $\mathbf{x} = (x_k)_{k\in[\![1;n]\!]}$. Notons $\mathscr{C} = \{\alpha_1, \alpha_2, \dots, \alpha_p\}$ les valeurs atteintes par le caractère, *i.e.* $x_k \in \mathscr{C}$.
L'ordre dans lequel on a recueilli les données ne présentant pas d'intérêt particulier, on a intérêt à regrouper les données par paquets. On appelle alors

- **effectif** de la valeur α_i, et on le note n_i, le nombre de fois que la valeur $\alpha_i \in \mathscr{C}$ est prise dans $\mathbf{x}$;
- **effectif cumulé** en α_i la somme $\sum_{j=1}^{i} n_j$;
- **fréquence** de la valeur α_i le rapport $f_i = \frac{n_i}{n}$;
- **fréquence cumulée** en α_i la somme $\sum_{j=1}^{i} f_j$.

Si on écrit la **série statistique** $\mathbf{x} = (x_k)_{k\in[\![1;n]\!]}$ comme $(\alpha_i, n_i)_{i\in[\![1;p]\!]}$ ou $(\alpha_i, f_i)_{i\in[\![1;p]\!]}$ on parle de **distribution statistique**.

EXEMPLE
Soit la série statistique $\mathbf{x} = (1,1,2,1,1,0,3,1)$.

- Elle contient $n = 8$ valeurs $x_1 = 1$, $x_2 = 1$, $x_3 = 2$, $x_4 = 1$, $x_5 = 1$, $x_6 = 0$, $x_7 = 3$, $x_8 = 1$;
- les valeurs atteintes sont $\mathscr{C} = \{\alpha_1 = 0, \alpha_2 = 1, \alpha_3 = 2, \alpha_4 = 3\}$;
- $n_1 = 1$, $n_2 = 5$, $n_3 = 1$, $n_4 = 1$;
- $f_1 = 1/8$, $f_2 = 5/8$, $f_3 = 1/8$, $f_4 = 1/8$;
- les effectifs cumulés sont respectivement 1, 6, 7, 8 ;
- les fréquences cumulées sont respectivement $1/8$, $6/8$, $7/8$, $8/8 = 1$.

DATA 8.1 (ENFANTS)
Considérons le nombre d'enfants par famille collectés dans un immeuble de $n = 80$ familles :

$$\mathbf{x} = (0,3,0,0,0,0,3,3,3,5,3,2,0,0,0,1,2,1,1,1,1,1,2,0,4,2,2,0,4,1,0,5,2,3,2,3,0,3,4,5,0,1,3,0,0,3,$$
$$1,0,0,0,2,0,0,0,1,0,3,4,4,0,0,0,0,1,5,2,0,3,2,0,1,0,2,4,0,1,3,3,0,5).$$

On a $\mathscr{C} = \{0,1,2,3,4,5\}$.
L'effectif n_i de chaque valeur $\alpha_i \in \mathscr{C}$ est le nombre d'observations de cette valeurs (*i.e.* combien de fois α_i apparaît dans $\mathbf{x}$). La fréquence f_i de la valeur α_i est le rapport de l'effectif n_i sur le nombre totale d'observations n :

Valeur α_i (Nombre d'enfants)	Effectif n_i (Nombre de familles)	Fréquence f_i (Proportion de familles)	Effectif cumulé	Fréquence cumulée
0	31	31/80	31	31/80
1	13	13/80	44	44/80
2	11	11/80	55	55/80
3	14	14/80	69	69/80
4	6	6/80	75	75/80
5	5	5/80	80	80/80
	$\sum_{i=1}^{p=6} n_i = 80$	$\sum_{i=1}^{p=6} f_i = 1$		

8.1.2. Série statistique et distribution statistique groupée

Lorsqu'un caractère comprend un grand nombre de valeurs, il est préférable de les regrouper. L'ensemble $\mathscr{C}$ des valeurs du caractère est alors partagé en intervalles disjoints $]\alpha_i;\alpha_{i+1}]$, appelés **classes**, avec $\alpha_i < \alpha_{i+1}$.
On appelle alors

- **amplitude** de la classe $]\alpha_i;\alpha_{i+1}]$ la largeur de l'intervalle ;
- **effectif** de la classe $]\alpha_i;\alpha_{i+1}]$, et on le note n_i, le nombre de valeurs de $\mathbf{x}$ qui appartiennent à cet interval (le nombre d'observations qui tombent dans cette classe) ;
- **effectif cumulé** en α_i le nombre de valeurs de $\mathbf{x}$ qui appartiennent à $]-\infty;\alpha_i]$;
- **fréquence** de la classe $]\alpha_i;\alpha_{i+1}]$ le rapport $f_i = \frac{n_i}{n}$;
- **fréquence cumulée** en α_i la somme $\sum_{j=1}^{i} f_j$.

Si on écrit la série statistique $\mathbf{x} = (x_k)_{k\in[\![1;n]\!]}$ comme $(]\alpha_i;\alpha_{i+1}], n_i)_{i\in[\![1;p]\!]}$ ou $(]\alpha_i;\alpha_{i+1}], f_i)_{i\in[\![1;p]\!]}$ on parle de **distribution statistique groupée**.
Le nombre de classes ne doit pas être trop grand pour que le nouveau tableau soit suffisamment clair, mais pas trop petit pour qu'il n'y ait pas de perte d'information trop importante. Il faut enfin que toutes les observations soient recouvertes par ces classes.

DATA 8.2 (AMPOULES)
Supposons qu'on ait recueilli les durée de vie (en heures) d'un lot d'ampoules :

$$\mathbf{x} = (2560, 229323551738, 2272, 2259, 2549, 1688, 2306, 2494, 2131, 1864, 2107, 2056, 2557, 1311, 2305, 2433, 2408, 1523,$$
$$2155, 2531, 2327, 1396, 2414, 2411, 2329, 1424, 2456, 2149, 2039, 1447, 1884, 2289, 2340, 1428, 2134, 2333, 1989, 1554,$$
$$2558, 2031, 2111, 1415, 2335, 2546, 2343, 1493, 2435, 2131, 2026, 1631, 2513, 2233, 2416, 1441, 2475, 2304, 2177, 1432,$$
$$1918, 2092, 2139, 1657, 2628, 2334, 2091, 1428, 2504, 2519, 2125, 1458, 2085, 2234, 2339, 1484, 2052, 2168, 2280, 1547,$$
$$2393, 2048, 1517, 1579, 2373, 2207, 1452, 1859, 2177, 2112, 1573, 1473, 2474, 2513, 1488, 1391, 2109, 2296, 1410, 1607,$$

2286, 2303, 1432, 1577, 2389, 1945, 1589, 1438, 2408, 1925, 1431, 1652, 2215, 2420, 1546, 1597, 2429, 2381, 1672, 1636).

On peut regrouper ces données en 8 classes de même amplitudes :

Durée de vie	Effectif	Fréquence
]1200;1400]	3	3/120
]1400;1600]	27	27/120
]1600;1800]	8	8/120
]1800;2000]	7	7/120
]2000;2200]	23	23/120
]2200;2400]	28	28/120
]2400;2600]	23	23/120
]2600;2800]	1	1/120
	$\sum_{i=1}^{p=8} n_i = 120$	$\sum_{i=1}^{p=8} f_i = 1$

On peut aussi subdiviser les trois classes de 2000 à 2600 en six classes et obtenir ainsi des classes d'amplitudes différentes :

Durée de vie	Effectif	Fréquence
]1200;1400]	3	3/120
]1400;1600]	27	27/120
]1600;1800]	8	8/120
]1800;2000]	7	7/120
]2000;2100]	9	9/120
]2100;2200]	14	14/120
]2200;2300]	11	11/120
]2300;2400]	17	17/120
]2400;2500]	13	13/120
]2500;2600]	10	10/120
]2600;2800]	1	1/120
	$\sum_{i=1}^{p=11} n_i = 120$	$\sum_{i=1}^{p=11} f_i = 1$

Sur un caractère qualitatif, le seul calcul numérique qu'on puisse effectuer est le dénombrement des unités statistiques dans chaque catégorie de la variable qualitative.

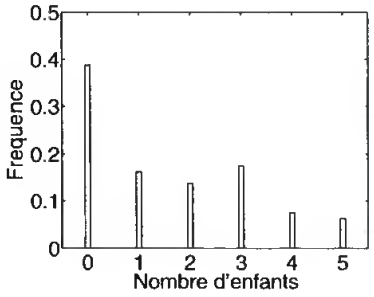
8.2. Données statistiques et leur représentation

8.2.1. Diagramme en bâtons

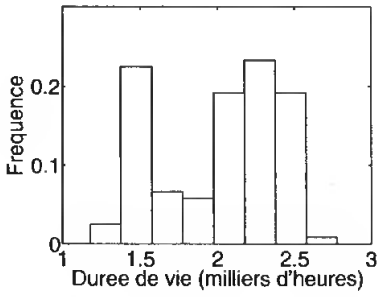
Dans le cas de données discrètes (ou d'une série statistique non groupée) on trace la plupart du temps un diagramme en bâtons des effectifs ou des fréquences. Des segments de droite verticaux sont dessinés. Chaque segment correspond à une classe (*i.e.* une modalité). La valeur de la classe est l'abscisse du segment, l'ordonnée de l'extrémité inférieure du segment est 0 et l'ordonnée de l'extrémité supérieure est l'effectif de la classe ou la fréquence. Les data 8.1 peut ainsi se représenter sous la forme du diagramme en bâtons donné sur la figure 8.1a.

8.2.2. Histogramme

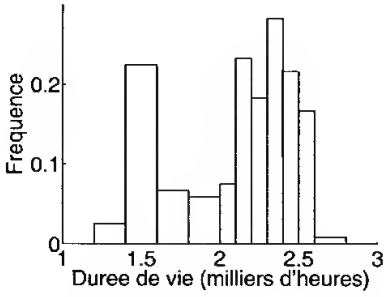
Dans le cas de données continues (ou d'une série statistique groupée), on regroupe d'abord les données par classes. On trace alors un histogramme constitué de rectangles verticaux. Les bases des rectangles sont déterminées par les classes. Les hauteurs de rectangles doivent être telles que les surfaces des rectangles sont proportionnelles aux effectifs des classes correspondantes.
Le travail est simple lorsque la largeur de chaque classe est la même : la hauteur d'un rectangle est alors prise égale à l'effectif (ou à la fréquence) de la classe correspondante. C'est le cas des data 8.2 avec la première subdivision en classes et on obtient alors la figure 8.1b.
Il arrive qu'on ait affaire à des classes non-régulières. Le tracé d'un histogramme doit alors prendre en compte la non-uniformité des largeurs des classes. Pour cela, on prend la plus petite des largeurs (ou amplitudes) comme largeur de référence, et multiplie la hauteur des rectangles par le rapport de leur largeur sur cette largeur minimale. C'est le cas de data 8.2 avec la deuxième subdivision en classes et on obtient alors la figure 8.1c.



(a) Distribution (diagramme en bâtons) du nombre d'enfants par famille.



(b) Histogramme des durées de vie des ampoules.



(c) Histogramme des durées de vie des ampoules.

FIGURE 8.1. – Exemples d'histogrammes.

8.3. Statistique descriptive univariée

Les tableaux et les diagrammes sont utiles, mais ils ne sont que des outils de visualisation. On cherche souvent, à partir de données quantitatives collectées, à extraire des caractéristiques chiffrées simples, des nombres qui révèlent les propriétés importantes de l'échantillon ou de la population. Nous nous intéressons à deux types de mesure : des mesures qui s'intéressent à la tendance centrale, *i.e.* à la plus représentative de toutes les données, et des mesures de la dispersion, *i.e.* combien les mesures de tendance centrale sont representatives de toutes les données.

8.3.1. Mesures de tendance centrale

Il s'agit de déterminer la valeur qui est la plus représentative de toutes les données.

Mode (ou classe modale)

Pour un caractère discret, le MODE est la valeur la plus fréquente que l'on trouve dans un échantillon.
Dans le cas de caractères continus ou plus généralement d'une série statistique groupée, on considère plutôt la CLASSE MODALE qui est le rapport fréquence/amplitude maximal. Le résultat dépend donc des classes choisies, ce qui rend cette notion peu pratique à utiliser.
Le mode n'est pas défini de manière unique. On peut trouver plusieurs classes avec le même effectif. On parle alors de distribution multi-modale.
Le mode est peu sensible aux valeurs extrêmes.

EXEMPLE

```
xx = [5 2 4 2 6]
mode(xx) % ans = 2

xx = [5 2 4 2 6 5]
mode(xx) % ans = 2
% If two, or more, values have the same frequency 'mode' returns the smallest
```

Médiane

La MÉDIANE est une valeur M telle qu'il y ait autant d'observations supérieures ou égales à M que d'observations inférieures ou égales à M. Pour calculer précisément la médiane, on commence par ordonner l'échantillon $\mathbf{x}$ par ordre croissant, et on note $\mathbf{y} = (y_k)_{k\in[\![1;n]\!]}$ l'échantillon ordonné tel que $y_1 \leq y_2 \leq \cdots \leq y_n$. Si l'échantillon comporte un nombre impair $2p+1$ d'observations, alors la médiane est

$$M(\mathbf{x}) = y_{p+1},$$

si l'échantillon est constitué d'un nombre pair $2p$ d'observations, alors la médiane est

$$M(\mathbf{x}) = \frac{y_p + y_{p+1}}{2}.$$

La médiane est peu sensible aux valeurs extrêmes et n'est pas forcément une modalité.

EXEMPLE
Si l'échantillon $\mathbf{x}$ est constitué de la suite d'entiers $(5,2,4,2,6)$, alors l'échantillon ordonné $\mathbf{y}$ est $(2,2,4,5,6)$. On a $n=5$ éléments, donc $p=\frac{n-1}{2}=2$ et $M(\mathbf{x})=y_{p+1}=y_3=4$.

```
xx = [5 2 4 2 6]
median(xx) % ans =4

xx = [5 2 4 2 6 30]
median(xx) % ans =4.5
```

Moyenne arithmétique

On peut définir la moyenne arithmétique d'une série statistique $\mathbf{x}=(x_k)_{k\in[\![1;n]\!]}$ comme étant le barycentre des données, affectées de coefficients égaux pour chaque individu :

$$\bar{\mathbf{x}}=\frac{1}{n}\sum_{k=1}^{n}x_k.$$

On appelle communément $\bar{\mathbf{x}}$ la moyenne de $\mathbf{x}$.
Si on écrit la série statistique $\mathbf{x}=(x_k)_{k\in[\![1;n]\!]}$ comme la distribution statistique $(\alpha_i,n_i)_{i\in[\![1;p]\!]}$ ou $(\alpha_i,f_i)_{i\in[\![1;p]\!]}$, alors

$$\bar{\mathbf{x}}=\sum_{j=1}^{p}f_j\alpha_j=\frac{1}{n}\sum_{j=1}^{p}n_j\alpha_j.$$

Remarque (Sensibilité aux valeurs extrêmes)
La moyenne est très sensible aux valeurs extrêmes. Par exemple, si on cherche la fortune moyenne des Français à partir d'un échantillon de 1000 personnes, si l'une d'entre elles possède un milliard d'euros, alors la fortune moyenne est supérieure à un million d'euros quelles que soient les fortunes des 999 autres, puisqu'elle vérifie :

$$\bar{\mathbf{x}}=\frac{1}{10^3}\sum_{k=1}^{10^3}x_k\geq\frac{1}{10^3}\left(\sum_{k=1}^{999}0+10^9\right)=10^6.$$

EXEMPLE

```
xx = [5 2 4 2 6]
mean(xx) % ans = 3.8

xx = [5 2 4 2 6 50]
mean(xx) % ans = 11.5
```

Dans le cas d'une distribution statistique groupée $(]\alpha_i;\alpha_{i+1}],n_i)_{i\in[\![1;p]\!]}$ dont on n'a pas toutes les données $\mathbf{x}$, il n'est pas possible de calculer la moyenne exactement. Si on ne dispose que du tableau des fréquences, alors on estime la moyenne par la formule

$$\bar{\mathbf{x}}\approx\sum_{i=1}^{p}f_i\frac{\alpha_i+\alpha_{i+1}}{2},$$

où $\frac{\alpha_i+\alpha_{i+1}}{2}$ est le centre de la i-ème classe et f_i sa fréquence.

Propriété 8.1 (Fusion de données)
Considérons la situation où on dispose de deux échantillons $\mathbf{u}=(u_k)_{k\in[\![1;n_1]\!]}$ et $\mathbf{v}=(v_k)_{k\in[\![1;n_2]\!]}$ de tailles n_1 et n_2 et de moyennes respectives $\bar{\mathbf{u}}$ et $\bar{\mathbf{v}}$. L'échantillon globale $\mathbf{x}$ fusion des deux échantillons $\mathbf{u}$ et $\mathbf{v}$ est de taille $n=n_1+n_2$ et sa moyenne est

$$\bar{\mathbf{x}}=\frac{n_1\bar{\mathbf{u}}+n_2\bar{\mathbf{v}}}{n}.$$

Autrement dit, lorsqu'on fusionne les résultats issus d'échantillons différents, on peut obtenir la moyenne de l'échantillon global sans avoir à refaire tous les calculs.
La médiane et le mode ne vérifient pas cette propriété.

Propriété 8.2 (Erreur quadratique)
Considérons la fonction $\mathscr{E} : \mathbb{R} \to \mathbb{R}_+$ définie par

$$\mathscr{E}(\mu) = \frac{1}{n}\sum_{k=1}^{n}(x_k - \mu)^2.$$

Elle atteint son minimum en $\mu = \bar{\mathbf{x}}$.

La fonction $\mathscr{E}$, appelée ERREUR QUADRATIQUE, est la moyenne des carrés des distances entre les x_k et le nombre réel μ. La moyenne $\bar{\mathbf{x}}$ est la constante qui approche au mieux les données au sens des moindres carrés.

PREUVE
$\mathscr{E}(\mu) \geq 0$ pour tout μ et

$$\mathscr{E}'(\mu) = -\frac{2}{n}\sum_{k=1}^{n}(x_k - \mu) = -2\left(\frac{1}{n}\sum_{k=1}^{n}(x_k - \frac{1}{n}n\mu\right) = -2\left(\frac{1}{n}\sum_{k=1}^{n}x_k - \mu\right) = -2(\bar{\mathbf{x}} - \mu).$$

Ainsi $\mu = \bar{\mathbf{x}}$ est un extremum et comme $\mathscr{E}$ est quadratique, il s'agit d'un minimum.

8.3.2. Mesures de dispersion

On vient d'examiner différentes mesures de la tendance centrale d'un échantillon. On va maintenant chercher une mesure de la variabilité d'un échantillon, c'est-à-dire un nombre qui est d'autant plus grand que les données de l'échantillon sont dispersées.

Variance

La dispersion d'un échantillon peut se visualiser en considérant les écarts à la moyenne, c'est-à-dire l'échantillon $\mathbf{v} = (v_k)_{k\in[\![1;n]\!]}$ avec $v_k = x_k - \bar{\mathbf{x}}$. On cherche à obtenir une valeur unique représentative de ces écarts. On ne va pas prendre la moyenne $\bar{\mathbf{v}}$ car elle est nulle quel que soit l'échantillon (d'après la linéarité de la moyenne arithmétique). On va donc prendre les carrés des écarts et calculer leur moyenne arithmétique. On obtient ainsi la VARIANCE de l'échantillon :

$$V(\mathbf{x}) = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{\mathbf{x}})^2.$$

Autrement dit, la variance est la valeur de l'erreur quadratique en son minimum (la moyenne minimise la fonction "erreur quadratique", la variance est l'évaluation de cette fonction dans le minimum) :

$$V(\mathbf{x}) = \mathscr{E}(\bar{\mathbf{x}}).$$

La variance est une quantité positive qui augmente avec la dispersion des données. Elle est nulle si et seulement si toutes les données sont égales.
Si on écrit la série statistique $\mathbf{x} = (x_k)_{k\in[\![1;n]\!]}$ comme la distribution statistique $(\alpha_i, n_i)_{i\in[\![1;p]\!]}$ ou $(\alpha_i, f_i)_{i\in[\![1;p]\!]}$, alors

$$V(\mathbf{x}) = \frac{1}{n}\sum_{j=1}^{p} n_j(\alpha_j - \bar{\mathbf{x}})^2 = \sum_{j=1}^{p} f_j(\alpha_j - \bar{\mathbf{x}})^2.$$

Théorème 8.3 (de Koenig, formule de Huygens)

$$V(\mathbf{x}) = \frac{\sum_{k=1}^{n} x_k^2}{n} - \bar{\mathbf{x}}^2.$$

Cette expression est utile pour calculer pratiquement la variance d'un échantillon donné.

PREUVE

$$V(\mathbf{x}) = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{\mathbf{x}})^2. = \frac{1}{n}\sum_{k=1}^{n}(x_k^2 - 2x_k\bar{\mathbf{x}} + \bar{\mathbf{x}}^2). = \frac{1}{n}\sum_{k=1}^{n}x_k^2 - 2\frac{1}{n}\sum_{k=1}^{n}x_k\bar{\mathbf{x}} + \frac{1}{n}\sum_{k=1}^{n}\bar{\mathbf{x}}^2 = \frac{1}{n}\sum_{k=1}^{n}x_k^2 - 2\bar{\mathbf{x}}^2 + \frac{1}{n}n\bar{\mathbf{x}}^2 = \frac{1}{n}\sum_{k=1}^{n}x_k^2 - \bar{\mathbf{x}}^2.$$

Si on écrit la série statistique $\mathbf{x} = (x_k)_{k\in[\![1;n]\!]}$ comme la distribution statistique $(\alpha_i, n_i)_{i\in[\![1;p]\!]}$ ou $(\alpha_i, f_i)_{i\in[\![1;p]\!]}$, alors la formule de Huygens devient

$$V(\mathbf{x}) = \frac{\sum_{j=1}^{p} n_j \alpha_j^2}{n} - \bar{\mathbf{x}}^2 = \sum_{j=1}^{p} f_j \alpha_j^2 - \bar{\mathbf{x}}^2.$$

Propriété 8.4 (Fusion de données)
Considérons la situation où on dispose de deux échantillons $\mathbf{u} = (u_k)_{k\in[\![1;n_1]\!]}$ et $\mathbf{v} = (v_k)_{k\in[\![1;n_2]\!]}$ de tailles n_1 et n_2 et de variances respectives $V(\mathbf{u})$ et $V(\mathbf{v})$. L'échantillon globale $\mathbf{x}$ fusion des deux échantillons $\mathbf{u}$ et $\mathbf{v}$ est de taille $n = n_1 + n_2$ et sa variance est

$$V(\mathbf{x}) = \frac{n_1}{n}\left(V(\mathbf{u}) + \bar{\mathbf{u}}^2\right) + \frac{n_2}{n}\left(V(\mathbf{v}) + \bar{\mathbf{v}}^2\right) - \left(\frac{n_1\bar{\mathbf{u}} + n_2\bar{\mathbf{v}}}{n}\right)^2.$$

PREUVE

$$V(\mathbf{x}) = \frac{1}{n}\sum_{k=1}^{n} x_k^2 - \bar{\mathbf{x}}^2 = \frac{1}{n}\sum_{k=1}^{n_1} u_k^2 + \frac{1}{n}\sum_{k=1}^{n_2} v_k^2 - \left(\frac{n_1\bar{\mathbf{u}} + n_2\bar{\mathbf{v}}}{n_1+n_2}\right)^2 = \frac{n_1}{n}\left(V(\mathbf{u}) + \bar{\mathbf{u}}^2\right) + \frac{n_2}{n}\left(V(\mathbf{v}) + \bar{\mathbf{v}}^2\right) - \left(\frac{n_1\bar{\mathbf{u}} + n_2\bar{\mathbf{v}}}{n}\right)^2.$$

Écart-type

La variance présente un inconvénient majeur : si les données s'expriment en unités physiques, la moyenne arithmétique s'exprime aussi dans cette unité, mais la variance s'exprime dans l'unité carrée. C'est pourquoi on a introduit la notion d'ÉCART-TYPE :

$$\sigma(\mathbf{x}) = \sqrt{V(\mathbf{x})}$$

Remarque (Diviser par n ou $n-1$?)
La variance est utilisée si la population est accessible dans sa totalité. Cependant d'ordinaire nous nous intéressons à une population dont on n'a pu mesurer qu'un échantillon. Dans ce cas, la meilleure estimation que l'on puisse faire de la variance de la population est

$$E(V) = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{\mathbf{x}})^2 = \frac{n}{n-1}V.$$

$E(V)$ est dit *variance corrigée* ou *estimateur non biaisé de la variance* de la population car si l'on multiplie le prélèvement d'échantillons de même effectif dans cette même population, la moyenne des $E(V)$ tend vers la variance réelle de la population, ce qui n'est pas le cas de V.
Même remarque pour l'*estimation de l'écart-type* de la population (ou *écart-type corrigé*)

$$s(\sigma) = \sqrt{E(V)} = \sqrt{\frac{n}{n-1}V} = \sqrt{\frac{n}{n-1}}\sigma$$

Faut-il calculer l'indice de dispersion de l'échantillon ou l'estimation de celui de la population, autrement dit, faut-il diviser par n ou $n-1$?

- Si l'on cherche à tirer des conclusions sur une population à partir d'un échantillon de celle-ci, $(n-1)$ sera utilisé dans les calculs,
- si l'on cherche à décrire un échantillon ou si la variable a été mesurée sur tous les individus de la population, c'est n qui sera utilisé.

EXEMPLE

```
xx = [5 2 4 2 6]
n=length(xx)

% Estimation de l'ecart-type
s=std(xx) % ans = 1.7889

% Estimation de la variance
V=var(xx) % ans = 3.2
% E=s*s

% Ecart-type
```

```
sigma=std(xx,1) % ans = 1.6
% sigma=sqrt((n-1)/n)*s

% Variance
var(xx,1) % ans = 2.56
% V=sigma*sigma
% V=(n-1)/n*E
```

EXEMPLE
Calculons le mode, la médiane, la moyenne arithmétique et la variance des data 8.1.

- **Mode** Le mode est 0.
- **Médiane** On a $n = 80$ éléments, donc $p = \frac{n}{2} = 40$ et $M(\mathbf{x}) = \frac{y_p + y_{p+1}}{2} = \frac{y_{40}+y_{41}}{2} = 1$ (car $y_i = 0$ pour $i = 1,\dots,31$, $y_i = 1$ pour $i = 32,\dots,44$ etc).
- **Moyenne** On a $\bar{\mathbf{x}} = \dfrac{0\times 31 + 1\times 13 + 2\times 11 + 3\times 14 + 4\times 6 + 5\times 5}{80} = 1.575$
- **Variance** On a $\dfrac{0^2\times 31 + 1^2\times 13 + 2^2\times 11 + 3^2\times 14 + 4^2\times 6 + 5^2\times 5}{80} = 5.05$ ainsi $V(\mathbf{x}) = 5.05 - 1.575 = 3.475$

```
clear all;

% Chargement des valeurs
load enfantsdata.dat ;
valeurs=sort(enfantsdata);

tt=0:5 ;
hist(valeurs,tt);
h = findobj(gca,'Type','patch');
set(h(1),'FaceColor','y','EdgeColor','k');

n=length(valeurs)
[effectif,c]=hist(valeurs,unique(valeurs))
freq=effectif/sum(effectif)

my_mode=valeurs(effectif==max(effectif))

my_moyenne=sum(c.*effectif)/sum(effectif)
my_moyenne=sum(c.*freq)
moyenne=mean(valeurs)

my_V=sum(valeurs.^2)/n-my_moyenne^2
V=var(valeurs,1)

my_sigma=sqrt(my_V)
sigma=std(valeurs,1)

mediane=median(valeurs)
figure()
pkg load statistics
boxplot(valeurs);
axis ([0,2,-1,6]);
```

Fractiles, quantiles

Les FRACTILES sont un autre moyen de quantifier la dispersion de données quantitatives. Le fractile à θ% d'un échantillon est la valeur qui sépare la fraction θ% des plus petites données de la fraction $(100-\theta)$% des plus grandes données.
Le fractile à 50% n'est autre que la médiane.
Les fractiles à 25%, 50% et 75% sont les trois quartiles.
Une mesure de la dispersion d'un échantillon est l'ESPACE INTER-QUARTILE qui est la différence entre le troisième quartile et le premier quartile ; c'est donc la largeur de l'intervalle qui contient l'échantillon duquel on a retiré les 25% plus grandes valeurs et les 25% plis petites valeurs. Qualitativement, plus l'espace inter-quartile est grand, plus la dispersion des données est grande. L'espace inter-quartile est moins sensible aux valeurs extrêmes que l'écart-type.

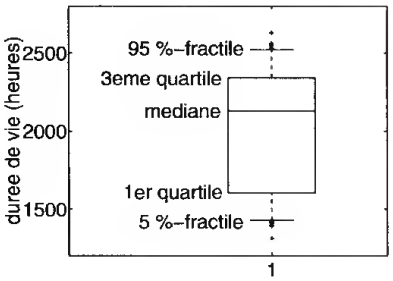


FIGURE 8.2. – Boîte à moustaches des durées de vie des ampoules (data 8.2).

8.3.3. Boîte à moustache

Une moyen très rapide de figurer le profil essentiel d'une série statistique quantitative est la boîte à moustaches (traduction française du terme "*Box and Whiskers Plot*" ou, en abrégé, "*Box Plot*"), aussi appelée boîte de distribution. Une telle boîte comprend

- ⋆ une échelle de valeurs sur l'axe vertical ;
- ⋆ le bord inférieur de la boîte correspond au premier quartile, noté Q_1 (*i.e.* le fractile à 25% ou quantile à 0.25) ;
- ⋆ le bord supérieur de la boîte correspond au troisième quartile, noté Q_3 (*i.e.* le fractile à 75% ou quantile à 0.75) ;
- ⋆ le trait horizontal au sein de la boîte correspond au deuxième quartile, noté Q_2 (*i.e.* la médiane) ;
- ⋆ les moustaches inférieure et supérieure, représentées par des traits verticaux de chaque coté de la boîte et qui se terminent par des traits horizontaux (il existe plusieurs façon de construire les moustaches, parfois elles correspondent aux fractiles à 5% et 95%, parfois au premier et neuvième décile, mais d'autres conventions existent) ;
- ⋆ les valeurs atypiques représentées par des cercles ou croix (on appelle ces données les outliers).

Une boîte avec des moustaches courtes indique que l'échantillon est assez dispersé.
Les boîtes à moustaches sont des résumés graphiques efficaces des données et sont donc très utiles pour comparer des distributions d'un groupe à l'autre. Contrairement à un histogramme, elle ne nécessitent pas de regrouper les observations en classes, ce qui est un avantage car le choix des classes est une opération subjective et qui influence fortement l'allure de l'histogramme construit à partir de celles-ci.
Sur la figure 8.2 on dessine la boîte à moustaches correspondant aux data 8.2. On constate que l'échantillon n'est pas équilibré, la médiane n'est pas vraiment au milieu des premier et troisième quartiles.

EXEMPLE
Utilisons les data 8.2.

```
clear all

% Chargement des valeurs
load mesures.dat
valeurs=sort(mesures);

% Nombre d'elements
n=length(valeurs)

%Moyenne
moyenne=mean(valeurs)

%Mediane
mediane=median(valeurs)

% Estimation de l'ecart-type sans biais
etsb=std(valeurs)

% Incertitude de type A
ua=etsb/sqrt(n)

% Definitions des bornes des intervalles
tt=1200:200:2800 ;
hist(valeurs,tt-100); # on passe le centre des intervalles
h = findobj(gca,'Type','patch');
display(h)
set(h(1),'FaceColor','r','EdgeColor','k');
```

```
% Decoupage en classes
ncl=length(tt)-1

%calcul des frequence
for i=1:ncl
  eff(i)=length(find(valeurs>tt(i) & valeurs<=tt(i+1))) ;
  fm(i)=eff(i)/n ;
end

% Borne gauche, borne droite, effectif, frequence
A=[tt(1:end-1)',tt(2:end)',eff',fm']

% Boite a moustache
figure()
pkg load statistics
boxplot(valeurs)
axis ([0,2]);

% The returned matrix has one column for each data set as follows:
% 1 Minimum = 1311.0
% 2 1st quartile = 1604.5
% 3 2nd quartile (median) = 2131.0
% 4 3rd quartile = 2340.8
% 5 Maximum = 2628.0
% 6 Lower confidence limit for median = 2025.5
% 7 Upper confidence limit for median = 2236.5
```

8.4. Statistique descriptive à deux caractères

Lorsque les observations portent simultanément sur deux caractères, on les présente sous la forme d'un tableau à double entrée. On définit alors la distribution conjointe, les distributions marginales et les distributions conditionnelles. L'étude de la distribution de deux variables se poursuit par celle de leur liaison. L'étude de la liaison entre les variables observées, appelée communément l'étude des corrélations, dépend de leur nature. Ici on n'envisagera que le cas de deux variables quantitatives non groupée en classes.

8.4.1. Distribution conjointe

Considérons donc une série statistique dont les observations portent sur deux caractères. On veut ici extraire des informations sur la distribution jointe des deux caractères et étudier leur dépendance. Désignons par

- $(\mathbf{x},\mathbf{y}) = (x_k, y_k)_{k\in[\![1;n]\!]}$ les n données brutes, généralement présentées sous la forme d'un tableau à deux colonnes ;
- $\mathscr{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_p\}$ les p modalités de $\mathbf{x}$, *i.e.* les p valeurs distinctes observées pour $\mathbf{x}$ (autrement dit $x_k \in \mathscr{A}$) ;
- $\mathscr{B} = \{\beta_1, \beta_2, \ldots, \beta_q\}$ les q modalités de $\mathbf{y}$, *i.e.* les q valeurs distinctes observées pour $\mathbf{y}$ (autrement dit $y_k \in \mathscr{B}$).

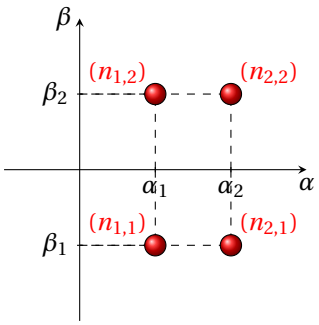
La répartition des n observations, ou **distribution conjointe**, suivant les modalités de $\mathbf{x}$ et $\mathbf{y}$ se présente sous forme d'un tableau à double entrée, appelée **tableau de contingence** :

Modalités de $\mathbf{y}$ / Modalités de $\mathbf{x}$	β_1	…	β_j	…	β_q	Effectif marginal de α_i
α_1	$n_{1,1}$	…	$n_{1,j}$	…	$n_{1,q}$	$n_{1,\cdot} = \sum_{j=1}^{q} n_{1,j}$
⋮	⋮		⋮		⋮	⋮
α_i	$n_{i,1}$	…	$n_{i,j}$	…	$n_{i,q}$	$n_{i,\cdot} = \sum_{j=1}^{q} n_{i,j}$
⋮	⋮		⋮		⋮	⋮
α_p	$n_{p,1}$	…	$n_{p,j}$	…	$n_{p,q}$	$n_{p,\cdot} = \sum_{j=1}^{q} n_{p,j}$
Effectif marginal de β_j	$n_{\cdot,1} = \sum_{i=1}^{p} n_{i,1}$	…	$n_{\cdot,j} = \sum_{i=1}^{p} n_{i,j}$	…	$n_{\cdot,q} = \sum_{i=1}^{p} n_{i,q}$	$n = \sum_{j=1}^{q} n_{\cdot,j} = \sum_{i=1}^{p} n_{i,\cdot}$

On appelle

- **effectif du couple** (α_i, β_j), et on le note $n_{i,j}$, le nombre de fois où le couple (α_i, β_j) est pris (*i.e.* le nombre de fois où la modalité α_i et la modalité β_j ont été observées simultanément) ;
- **fréquence du couple** (α_i, β_j) le rapport $f_{i,j} = \frac{n_{i,j}}{n}$.

Si on écrit la série statistique $(x_k, y_k)_{k\in[\![1;n]\!]}$ comme $((\alpha_i, \beta_j), n_{i,j})_{\substack{i\in[\![1;p]\!]\\ j\in[\![1;q]\!]}}$ ou $((\alpha_i, \beta_j), f_{i,j})_{\substack{i\in[\![1;p]\!]\\ j\in[\![1;q]\!]}}$ on parle de distribution conjointe.
On peut bien sûr représenter la série statistique ou la distribution conjointe sur un plan comme un nuage de points : chaque point correspond à un couple (α_i, β_j) affecté de son poids $n_{i,j}$, autrement dit chaque point correspond à une observation (x_k, y_k) et à coté on indique combien de fois cette observation apparaît. Il y aura donc $p \times q$ points (autant que de cases que dans le tableau de contingence), chaque point se trouvant sur un coin de la grille de coordonnées (α_i, β_j). Si pour un couple on a $n_{i,j} = 0$, on n'affichera pas de point. Si $n_{i,j} = 1$ pour tout $i = 1, \dots p$ et $j = 1, \dots q$, on a le nuage de points classique vu au chapitre précédent.



8.4.2. Distributions marginales

On peut bien sûr mener une étude statistique de chacun des caractères séparément, *i.e.* calculer la moyenne et la variance de chacune des séries simples ($\bar{\mathbf{x}}$, $\bar{\mathbf{y}}$, $V(\mathbf{x})$, $V(\mathbf{y})$). On appelle

- **effectif marginal** de α_i, et on le note $n_{i,\cdot}$, le nombre total d'observations de la modalité α_i de $\mathbf{x}$ quelle que soit la modalité de $\mathbf{y}$:

$$n_{i,\cdot} = \sum_{j=1}^{q} n_{i,j};$$

- **effectif marginal** de β_j, et on le note $n_{\cdot,j}$, total d'observations de la modalité β_j de $\mathbf{y}$ quelle que soit la modalité de $\mathbf{x}$:

$$n_{\cdot,j} = \sum_{i=1}^{p} n_{i,j};$$

- **fréquence marginale** de α_i le rapport $f_{i,\cdot} = \frac{n_{i,\cdot}}{n} = \sum_{j=1}^{q} f_{i,j}$;
- **fréquence marginale** de β_j le rapport $f_{\cdot,j} = \frac{n_{\cdot,j}}{n} = \sum_{i=1}^{p} f_{i,j}$.

On a bien évidemment

$$\sum_{i=1}^{p} n_{i,\cdot} = \sum_{j=1}^{q} n_{\cdot,j} = n \qquad \sum_{i=1}^{p} f_{i,\cdot} = \sum_{j=1}^{q} f_{\cdot,j} = 1.$$

Si on écrit la série statistique $\mathbf{x}$ comme $(\alpha_i, n_{i,\cdot})_{i\in[\![1;p]\!]}$ ou $(\alpha_i, f_{i,\cdot})_{i\in[\![1;p]\!]}$ on parle de distribution marginale de $\mathbf{x}$; de la même manière si on écrit la série statistique $\mathbf{y}$ comme $(\beta_j, n_{\cdot,j})_{j\in[\![1;q]\!]}$ ou $(\beta_j, f_{\cdot,j})_{j\in[\![1;q]\!]}$ on parle de distribution marginale de $\mathbf{y}$.
Disposant d'une distribution conjointe, on peut déduire les distributions marginales qui permettent d'étudier séparément chaque variable en représentant graphiquement sa distribution et s'il s'agit d'une variable quantitative, en calculant ses caractéristiques de tendance centrale, de dispersion, de forme…
On appelle

- **moyenne marginale** de $\mathbf{x}$ la quantité

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{p} n_{i,\cdot}\alpha_i = \sum_{i=1}^{p} f_{i,\cdot}\alpha_i$$

- **moyenne marginale** de $\mathbf{y}$ la quantité

$$\bar{\mathbf{y}} = \frac{1}{n}\sum_{j=1}^{q} n_{\cdot,j}\beta_j = \sum_{j=1}^{q} f_{\cdot,j}\beta_j$$

• **variance marginale** de **x** la quantité

$$V(\mathbf{x}) = \frac{\sum_{k=1}^{n}(x_k - \bar{\mathbf{x}})^2}{n} = \frac{\sum_{k=1}^{n} x_k^2}{n} - \bar{\mathbf{x}}^2 = \frac{1}{n}\sum_{i=1}^{p} n_{i,\cdot}(\alpha_i)^2 - \bar{\mathbf{x}}^2 = \sum_{i=1}^{p} f_{i,\cdot}\alpha_i^2 - \bar{\mathbf{x}}^2$$

• **variance marginale** de **y** la quantité

$$V(\mathbf{y}) = \frac{\sum_{k=1}^{n}(y_k - \bar{\mathbf{y}})^2}{n} = \frac{\sum_{k=1}^{n} y_k^2}{n} - \bar{\mathbf{y}}^2 = \frac{1}{n}\sum_{j=1}^{q} n_{\cdot,j}(\beta_j)^2 - \bar{\mathbf{y}}^2 = \sum_{j=1}^{q} f_{\cdot,j}\beta_j^2 - \bar{\mathbf{y}}^2.$$

8.4.3. Distributions conditionnelles

Une distribution à deux caractères présente deux types de distributions conditionnelles : les distributions conditionnelles de **x** selon **y** et les distributions conditionnelles de **y** selon **x**.

• **Distributions conditionnelles de y selon x**

Considérons la sous-population correspondante aux individus tels que $x = \alpha_i$.

La distribution de la variable **y** sachant $x = \alpha_i$ est appelée distribution conditionnelle de **y** pour $x = \alpha_i$. Il existe p distributions conditionnelle de **y** sachant $x = \alpha_i$, car $i = 1, \ldots, p$.

Modalités de **y** sachant α_i	β_1	…	β_j	…	β_q	Effectif marginal de α_i
α_i	$n_{i,1}$	…	$n_{i,j}$	…	$n_{i,q}$	$n_{i,\cdot} = \sum_{j=1}^{q} n_{i,j}$

Chaque distribution contient $n_{i,\cdot}$ observations et on peut calculer les quantités conditionnelles suivantes :

• **fréquence conditionnelle** de β_j sachant α_i comme la quantité

$$f_{j|i} = \frac{n_{i,j}}{n_{i,\cdot}} = \frac{f_{i,j}}{f_{i,\cdot}} \text{ avec } \sum_{j=1}^{q} f_{j|i} = 1;$$

• **distribution conditionnelle des fréquences** de **y** sachant α_i la distribution $(\beta_j, f_{j|i})_{j\in[\![1;q]\!]}$;

• **moyenne conditionnelle** de **y** sachant α_i la quantité

$$\bar{\mathbf{y}}\big|_i = \frac{1}{n_{i,\cdot}}\sum_{j=1}^{q} n_{i,j}\beta_j = \sum_{j=1}^{q} f_{j|i}\beta_j;$$

• **variance conditionnelle** de **y** sachant α_i la quantité

$$V_i(\mathbf{y}) = \frac{1}{n_{i,\cdot}}\sum_{j=1}^{q} n_{i,j}\beta_j^2 - \bar{\mathbf{y}}\big|_i^2 = \sum_{j=1}^{q} f_{j|i}\beta_j^2 - \bar{\mathbf{y}}\big|_i^2.$$

Les p modalités de **x** induisant une partition des n observations en p sous-groupes, la moyenne $\bar{\mathbf{y}}$ peut s'exprimer comme somme pondérées des p moyennes $\bar{\mathbf{y}}\big|_i$:

$$\bar{\mathbf{y}} = \sum_{i=1}^{p} \bar{\mathbf{y}}\big|_i f_{i,\cdot}$$

Il est fréquent de présenter les fréquences conditionnelles $f_{j|i}$ de **y** dans un tableau dont toutes les sommes en ligne

sont égales à 1 ; ce tableau est appelé tableau des profils en ligne :

Modalités de **y** / Modalités de **x**	β_1	…	β_j	…	β_q	
α_1	$f_{1\|1}$	…	$f_{j\|1}$	…	$f_{q\|1}$	1
⋮	⋮		⋮		⋮	⋮
α_i	$f_{1\|i}$	…	$f_{j\|i}$	…	$f_{q\|i}$	1
⋮	⋮		⋮		⋮	⋮
α_p	$f_{1\|p}$	…	$f_{j\|p}$	…	$f_{q\|p}$	1
Fréquence marginale de β_j	$f_{\cdot,1}$	…	$f_{\cdot,j}$	…	$f_{\cdot,q}$	1

• **Distributions conditionnelles de x selon y**

De manière analogue, considérons maintenant la sous-population correspondante aux individus tels que $y = \beta_j$. La distribution de la variable **x** sachant $y = \beta_j$ est appelée distribution conditionnelle de **x** pour $y = \beta_j$. Il existe q distributions conditionnelle de **x** sachant $y = \beta_j$ car $j = 1, \ldots, q$.

Modalités de **x** sachant β_j	β_j
α_1	$n_{1,j}$
⋮	⋮
α_i	$n_{i,j}$
⋮	⋮
α_p	$n_{p,j}$
Effectif marginal de β_j	$n_{\cdot,j} = \sum_{i=1}^{p} n_{i,j}$

Chaque distribution contient $n_{\cdot,j}$ observations et on peut définir les quantités conditionnelles suivantes :

• **fréquence conditionnelle** de α_i sachant β_j comme la quantité

$$f_{i|j} = \frac{n_{i,j}}{n_{\cdot,j}} = \frac{f_{i,j}}{f_{\cdot,j}}$$

• **distribution conditionnelle des fréquences** de **x** sachant β_j la distribution $(\alpha_i, f_{i|j})_{i \in [\![1;p]\!]}$;

• **moyenne conditionnelle** de **x** sachant β_j la quantité

$$\bar{\mathbf{x}}|_j = \frac{1}{n_{\cdot,j}} \sum_{i=1}^{p} n_{i,j} \alpha_i = \sum_{i=1}^{p} f_{j|i} \alpha_i$$

• **variance conditionnelle** de **x** sachant β_j la quantité

$$V_j(\mathbf{x}) = \frac{1}{n_{\cdot,j}} \sum_{i=1}^{p} n_{i,j} \alpha_i^2 - \bar{\mathbf{x}}|_j^2 = \sum_{i=1}^{p} f_{i|j} \alpha_i^2 - \bar{\mathbf{x}}|_j^2 .$$

Les q modalités de **y** induisant une partition des n observations en q sous-groupes, la moyenne $\bar{\mathbf{x}}$ peut s'exprimer comme somme pondérées des q moyennes $\bar{\mathbf{x}}|_j$:

$$\bar{\mathbf{x}} = \sum_{j=1}^{q} \bar{\mathbf{x}}|_j \, f_{\cdot,j}$$

De même on présente les fréquences conditionnelles $f_{i|j}$ de **x** dans un tableau dont toutes les sommes en colonne

sont égales à 1 ; ce tableau est appelé tableau des profils en colonne :

Modalités de **y** / Modalités de **x**	β_1	…	β_j	…	β_q	Fréquence marginale de α_i
α_1	$f_{1\|1}$	…	$f_{1\|j}$	…	$f_{1\|q}$	$f_{1,\cdot}$
⋮	⋮		⋮		⋮	⋮
α_i	$f_{i\|1}$	…	$f_{i\|j}$	…	$f_{i\|q}$	$f_{i,\cdot}$
⋮	⋮		⋮		⋮	⋮
α_p	$f_{p\|1}$	…	$f_{p\|j}$	…	$f_{p\|q}$	$f_{p,\cdot}$
	1	…	1	…	1	1

EXEMPLE

Soient les données brutes $\big((1,0),(1,2),(2,0),(2,2),(2,2),(1,1)\big)$, alors $n=6$.

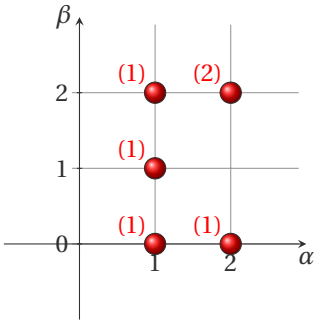
On a $x_k \in \mathscr{A} = \{1,2\}$ et $y_k \in \mathscr{B} = \{0,1,2\}$ pour tout $k = 1,2,\ldots,n$, ainsi $p=2$ et $q=3$. Écrivons les observations dans un tableau à deux colonnes :

x	**y**
1	0
1	2
2	0
2	2
2	2
1	1

• **Distribution conjointe et distributions marginales**

Le tableau des contingences avec les effectifs de chaque couple et les effectifs marginaux est

$\mathscr{A}$ / $\mathscr{B}$	$\beta_1=0$	$\beta_2=1$	$\beta_3=2$	Effectif marginal de α_i
$\alpha_1=1$	$n_{1,1}=1$	$n_{1,2}=1$	$n_{1,3}=1$	$n_{1,\cdot}=3$
$\alpha_2=2$	$n_{2,1}=1$	$n_{2,2}=0$	$n_{2,3}=2$	$n_{2,\cdot}=3$
Effectif marginal de β_j	$n_{\cdot,1}=2$	$n_{\cdot,2}=1$	$n_{\cdot,3}=3$	$n=6$



Le tableau des contingences avec les fréquences de chaque couple et les fréquences marginales est

$\mathscr{A}$ / $\mathscr{B}$	$\beta_1=0$	$\beta_2=1$	$\beta_3=2$	Fréquence marginale de α_i
$\alpha_1=1$	$f_{1,1}=1/6$	$f_{1,2}=1/6$	$f_{1,3}=1/6$	$f_{1,\cdot}=3/6$
$\alpha_2=2$	$f_{2,1}=1/6$	$f_{2,2}=0/6$	$f_{2,3}=2/6$	$f_{2,\cdot}=3/6$
Fréquence marginale de β_j	$f_{\cdot,1}=2/6$	$f_{\cdot,2}=1/6$	$f_{\cdot,3}=3/6$	1

Les moyennes marginales de **x** et **y** sont

$$\bar{\mathbf{x}} = \frac{1}{n}\left(n_{1,\cdot}\alpha_1 + n_{2,\cdot}\alpha_2\right) = \frac{1}{6}\left(3\alpha_1 + 3\alpha_2\right) = \frac{3}{2},$$

$$\bar{\mathbf{y}} = \frac{1}{n}\left(n_{\cdot,1}\beta_1 + n_{\cdot,2}\beta_2 + n_{\cdot,3}\beta_3\right) = \frac{1}{6}\left(2\beta_1 + 1\beta_2 + 3\beta_3\right) = \frac{7}{6}.$$

• **Distributions conditionnelles y sachant x**

• **y sachant α_1** On ne considère que la ligne de la modalité α_1 :

$\mathscr{A}$ \ $\mathscr{B}$	$\beta_1 = 0$	$\beta_2 = 1$	$\beta_3 = 2$	Fréquence marginale de α_1
$\alpha_1 = 1$	1/6	1/6	1/6	$f_{1,\cdot} = 3/6$

$$f_{j=1|i=1} = \frac{f_{i,j}}{f_{i,\cdot}} = \frac{1/6}{3/6} = \frac{1}{3} \qquad \text{fréquence conditionnelle de } \beta_1 \text{ sachant } \alpha_1$$
$$f_{j=2|i=1} = \frac{f_{i,j}}{f_{i,\cdot}} = \frac{1/6}{3/6} = \frac{1}{3} \qquad \text{fréquence conditionnelle de } \beta_2 \text{ sachant } \alpha_1$$
$$f_{j=3|i=1} = \frac{f_{i,j}}{f_{i,\cdot}} = \frac{1/6}{3/6} = \frac{1}{3} \qquad \text{fréquence conditionnelle de } \beta_3 \text{ sachant } \alpha_1$$

De plus,

$$\sum_{j=1}^{q=3} f_{j|i=1} = 1$$
$$\bar{\mathbf{y}}\big|_{i=1} = \sum_{j=1}^{q=3} f_{j|i=1}\beta_j = \frac{1}{3}\beta_1 + \frac{1}{3}\beta_2 + \frac{1}{3}\beta_3 = 1 \qquad \text{moyenne conditionnelle de } \mathbf{y} \text{ sachant } \alpha_1$$

• **y sachant α_2** On ne considère que la ligne de la modalité α_2 :

$\mathscr{A}$ \ $\mathscr{B}$	$\beta_1 = 0$	$\beta_2 = 1$	$\beta_3 = 2$	Fréquence marginale de α_2
$\alpha_2 = 2$	1/6	0/6	2/6	$f_{2,\cdot} = 3/6$

$$f_{j=1|i=2} = \frac{f_{i,j}}{f_{i,\cdot}} = \frac{1/6}{3/6} = \frac{1}{3} \qquad \text{fréquence conditionnelle de } \beta_1 \text{ sachant } \alpha_2$$
$$f_{j=2|i=2} = \frac{f_{i,j}}{f_{i,\cdot}} = \frac{0/6}{3/6} = 0 \qquad \text{fréquence conditionnelle de } \beta_2 \text{ sachant } \alpha_2$$
$$f_{j=3|i=2} = \frac{f_{i,j}}{f_{i,\cdot}} = \frac{2/6}{3/6} = \frac{2}{3} \qquad \text{fréquence conditionnelle de } \beta_3 \text{ sachant } \alpha_2$$

De plus,

$$\sum_{j=1}^{q=3} f_{j|i=2} = 1$$
$$\bar{\mathbf{y}}\big|_{i=2} = \sum_{j=1}^{q=3} f_{j|i=2}\beta_j = \frac{1}{3}\beta_1 + 0\beta_2 + \frac{2}{3}\beta_3 = \frac{4}{3} \qquad \text{moyenne conditionnelle de } \mathbf{y} \text{ sachant } \alpha_2$$

• **tableau des profils en ligne** $f_{j|i}$

Modalités de **y** / Modalités de **x**	$\beta_1 = 0$	$\beta_2 = 1$	$\beta_3 = 2$	
$\alpha_1 = 1$	$f_{1\|1} = 1/3$	$f_{2\|1} = 1/3$	$f_{3\|1} = 1/3$	1
$\alpha_2 = 2$	$f_{1\|2} = 1/3$	$f_{2\|2} = 0$	$f_{3\|2} = 2/3$	1
Fréquence marginale de β_j	$f_{\cdot,1} = 2/6$	$f_{\cdot,2} = 1/6$	$f_{\cdot,3} = 3/6$	1

On a bien

$$\sum_{i=1}^{p=2} \bar{\mathbf{y}}|_i \, f_{i,\cdot} = 1\frac{3}{6} + \frac{4}{3}\frac{3}{6} = \frac{7}{6} = \bar{\mathbf{y}},$$

• **Distributions conditionnelles x sachant y**

• **x sachant β_1** On ne considère que la colonne de la modalité β_1 :

$\mathscr{A}$ \ $\mathscr{B}$	$\beta_1 = 0$
$\alpha_1 = 1$	1/6
$\alpha_2 = 2$	1/6
Fréquence marginale de β_1	$f_{\cdot,1} = 2/6$

$$f_{i=1|j=1} = \frac{f_{i,j}}{f_{\cdot,j}} = \frac{1/6}{2/6} = \frac{1}{2} \qquad \text{fréquence conditionnelle de } \alpha_1 \text{ sachant } \beta_1$$

$$f_{i=2|j=1} = \frac{f_{i,j}}{f_{\cdot,j}} = \frac{1/6}{2/6} = \frac{1}{2} \qquad \text{fréquence conditionnelle de } \alpha_2 \text{ sachant } \beta_1$$

De plus,

$$\sum_{i=1}^{p=2} f_{i|j=1} = 1$$

$$\bar{\mathbf{x}}|_{j=1} = \sum_{i=1}^{p=2} f_{i|j=1}\alpha_i = \frac{1}{2}\alpha_1 + \frac{1}{2}\alpha_2 = \frac{3}{2} \qquad \text{moyenne conditionnelle de } \mathbf{x} \text{ sachant } \beta_1$$

• **x sachant β_2** On ne considère que la colonne de la modalité β_2 :

$\mathscr{A}$ \ $\mathscr{B}$	$\beta_2 = 1$
$\alpha_1 = 1$	1/6
$\alpha_2 = 2$	0/6
Fréquence marginale de β_2	$f_{\cdot,2} = 1/6$

$$f_{i=1|j=2} = \frac{f_{i,j}}{f_{\cdot,j}} = \frac{1/6}{1/6} = 1 \qquad \text{fréquence conditionnelle de } \alpha_1 \text{ sachant } \beta_2$$

$$f_{i=2|j=2} = \frac{f_{i,j}}{f_{\cdot,j}} = \frac{0/6}{1/6} = 0 \qquad \text{fréquence conditionnelle de } \alpha_2 \text{ sachant } \beta_2$$

De plus,

$$\sum_{i=1}^{p=2} f_{i|j=2} = 1$$

$$\bar{\mathbf{x}}|_{j=2} = \sum_{i=1}^{p=2} f_{i|j=2}\alpha_i = 1\alpha_1 + 0\alpha_2 = 1 \qquad \text{moyenne conditionnelle de } \mathbf{x} \text{ sachant } \beta_2$$

• **x sachant β_3** On ne considère que la colonne de la modalité β_3 :

$\mathscr{A}$ \ $\mathscr{B}$	$\beta_3 = 2$
$\alpha_1 = 1$	1/6
$\alpha_2 = 2$	2/6
Fréquence marginale de β_3	$f_{\cdot,3} = 3/6$

$$f_{i=1|j=3} = \frac{f_{i,j}}{f_{\cdot,j}} = \frac{1/6}{3/6} = \frac{1}{3} \qquad \text{fréquence conditionnelle de } \alpha_1 \text{ sachant } \beta_3$$

$$f_{i=2|j=3} = \frac{f_{i,j}}{f_{\cdot,j}} = \frac{2/6}{3/6} = \frac{2}{3} \qquad \text{fréquence conditionnelle de } \alpha_2 \text{ sachant } \beta_3$$

De plus,

$$\sum_{i=1}^{p=3} f_{i|j=3} = 1$$

$$\bar{\mathbf{x}}|_{j=3} = \sum_{i=1}^{p=2} f_{i|j=3}\alpha_i = \frac{1}{3}\alpha_1 + \frac{2}{3}\alpha_2 = \frac{5}{3} \qquad \text{moyenne conditionnelle de } \mathbf{x} \text{ sachant } \beta_3.$$

- **tableau des profils en colonne** $f_{i|j}$

Modalités de **y** / Modalités de **x**	$\beta_1 = 0$	$\beta_2 = 1$	$\beta_3 = 2$	Fréquence marginale de α_i
$\alpha_1 = 1$	$f_{1\|1} = 1/2$	$f_{1\|2} = 1$	$f_{1\|3} = 1/3$	$f_{1,\cdot}$
$\alpha_2 = 2$	$f_{2\|1} = 1/2$	$f_{2\|2} = 0$	$f_{2\|3} = 2/3$	$f_{p,\cdot}$
	1	1	1	1

On a bien

$$\sum_{j=1}^{q=3} \bar{\mathbf{x}}|_j \, f_{\cdot,j} = \frac{3}{2}\frac{2}{6} + 1\frac{1}{6} + \frac{5}{3}\frac{3}{6} = \frac{3}{2} = \bar{\mathbf{x}}.$$

8.4.4. Indépendance statistique

Si tous les profils en colonne du tableau en colonne sont identiques, cela signifie que la distribution de la variable **x** ne dépend pas de la variable **y**, on dit alors que les variables **x** et **y** sont statistiquement indépendantes dans l'ensemble des n individus considérés, et dans ce cas toutes les distributions conditionnelles de **x** sont identiques à la distribution marginale de **x**. Par raison de symétrie, l'indépendance statistique entre **x** et **y** implique aussi des profils en ligne identiques à la distribution marginale de **y**.
Les deux séries **x** et **y** sont indépendantes si et seulement si

$$\begin{cases} f_{i|j} = f_{i,\cdot}, & \textit{i.e.} \text{ la distribution conditionnelle des fréquences de } \alpha_i \text{ sachant } \beta_j \text{ ne dépend pas de } j \\ f_{j|i} = f_{\cdot,j}, & \textit{i.e.} \text{ la distribution conditionnelle des fréquences de } \beta_j \text{ sachant } \alpha_i \text{ ne dépend pas de } i. \end{cases}$$

De plus, si les deux séries sont indépendantes, alors pour tout $i = 1,\dots,p$ et $j = 1,\dots,q$

$$f_{i,j} = f_{i,\cdot} f_{\cdot,j}$$

Lorsque deux variables dépendent statistiquement l'une de l'autre, on cherche à évaluer l'intensité de leur liaison et dans le cas de deux variables quantitatives, on examine si on peut les considérer liées par une relation linéaire.

8.4.5. Liaison entre deux variables quantitatives : covariance et corrélation

La dispersion d'une série bivarié $(x_k, y_k)_{k\in[\![1;n]\!]}$ peut se visualiser en considérant les écarts aux deux moyennes. On cherche à obtenir une valeur unique représentative de ces écarts. On obtient ainsi la **covariance** de la série $(x_k, y_k)_{k\in[\![1;n]\!]}$:

$$C(\mathbf{x},\mathbf{y}) = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{\mathbf{x}})(y_k - \bar{\mathbf{y}}).$$

Si on écrit la série comme la distribution $((\alpha_i,\beta_j), n_{i,j})_{\substack{i\in[\![1;p]\!]\\ j\in[\![1;q]\!]}}$ ou $((\alpha_i,\beta_j), f_{i,j})_{\substack{i\in[\![1;p]\!]\\ j\in[\![1;q]\!]}}$, on a

$$C(\mathbf{x},\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{i,j}(\alpha_i - \bar{\mathbf{x}})(\beta_j - \bar{\mathbf{y}}) = \sum_{j=1}^{q} f_{i,j}(\alpha_i - \bar{\mathbf{x}})(\beta_j - \bar{\mathbf{y}}).$$

Propriété 8.5

1. $C(\mathbf{x},\mathbf{y}) = C(\mathbf{y},\mathbf{x})$
2. $C(\mathbf{x},\mathbf{x}) = V(\mathbf{x})$ et $C(\mathbf{y},\mathbf{y}) = V(\mathbf{y})$
3. $V(\mathbf{x}+\mathbf{y}) = V(\mathbf{x}) + 2C(\mathbf{x},\mathbf{y}) + V(\mathbf{y})$
4. $C(a\mathbf{x}+b, c\mathbf{y}+d) = acC(\mathbf{x},\mathbf{y})$ pour tout $a,b,c,\in\mathbb{R}$
5. $C(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{p}\sum_{j=1}^{q}\alpha_i\beta_j f_{ij} - \bar{\mathbf{x}}\bar{\mathbf{y}}$
6. $|C(\mathbf{x},\mathbf{y})| = \sqrt{V(\mathbf{x})V(\mathbf{y})}$

Si **x** et **y** sont indépendantes alors la covariance est nulle. La réciproque est fausse : en effet la covariance mesure uniquement la dépendance linéaire.

Remarque (Diviser par n ou $n-1$?)
Dans la définition ci-dessus, le dénominateur est n. Si l'on tente d'estimer la covariance de la population à partir d'un échantillon il faudra diviser par $(n-1)$. Les notations de la covariance de l'échantillon et de l'estimation de celle de la population ne sont pas en générale distinguables. Ainsi, lorsqu'on utilise un logiciel, toujours faire un calcul d'essai pour connaître la formule utilisée. Dans Octave ou Matlab c'est $(n-1)$ qui est utilisé par défaut, mais on peut forcer l'utilisation de n, comme on voit dans l'exemple ci-dessous.
Dans ce chapitre on utilisera la notation

$$E(C(\mathbf{x},\mathbf{y})) = \frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{\mathbf{x}})(y_k-\bar{\mathbf{y}}) = \frac{n-1}{n}C(\mathbf{x},\mathbf{y}).$$

Comme pour la variance, on dispose d'une formule alternative pour la covariance qu'on utilise en pratique pour calculer une covariance :

Propriété 8.6

$$C(\mathbf{x},\mathbf{y}) = \frac{\sum_{k=1}^{n}x_k y_k}{n} - \bar{\mathbf{x}}\bar{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q}n_{ij}\alpha_i\beta_j - \bar{\mathbf{x}}\bar{\mathbf{y}}.$$

EXEMPLE
Considérons l'échantillon bivarié $((1,1),(2,3),(3,5))$. On a

$$\mathbf{x} = (1,2,3) \qquad \mathbf{y} = (1,3,5) \qquad \bar{\mathbf{x}} = 2 \qquad \bar{\mathbf{y}} = 3$$

ainsi

$$C(\mathbf{x},\mathbf{y}) = \frac{1}{n}\sum_{k=1}^{n}(x_k-\bar{\mathbf{x}})(y_k-\bar{\mathbf{y}}) = \frac{(1-2)(1-3)+(2-2)(3-3)+(3-2)(5-3)}{3} = \frac{4}{3}$$

tandis que

$$E(C(\mathbf{x},\mathbf{y})) = \frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{\mathbf{x}})(y_k-\bar{\mathbf{y}}) = \frac{(1-2)(1-3)+(2-2)(3-3)+(3-2)(5-3)}{2} = \frac{4}{2} = 2.$$

'Octave'

```
x = [1 2 3];
y = [1 3 5];
E_cov = cov(x,y) % ans = 2
Cov = cov(x,y,1) % ans = 1.3333
% notre covariance
n=length(x)
moy_x = mean(x)
moy_y = mean(y)
my_cov = sum( (x-moy_x).*(y-moy_y))/n % ans = 1.3333
```

La covariance joue un rôle analogue à la variance dans le cas de deux caractères : elle mesure la dispersion conjointe des deux caractères. La corrélation joue un rôle analogue à l'écart type.

En supposant $V(\mathbf{x}) > 0$ et $V(\mathbf{y}) > 0$, c'est-à-dire que $n \geq 2$ et les x_k (resp. les y_k) ne sont pas tous égaux, on peut définir le **coefficient de corrélation linéaire (de Bravais-Pearson)** :

$$r(\mathbf{x},\mathbf{y}) = \frac{C(\mathbf{x},\mathbf{y})}{\sqrt{V(\mathbf{x})V(\mathbf{y})}}.$$

On a

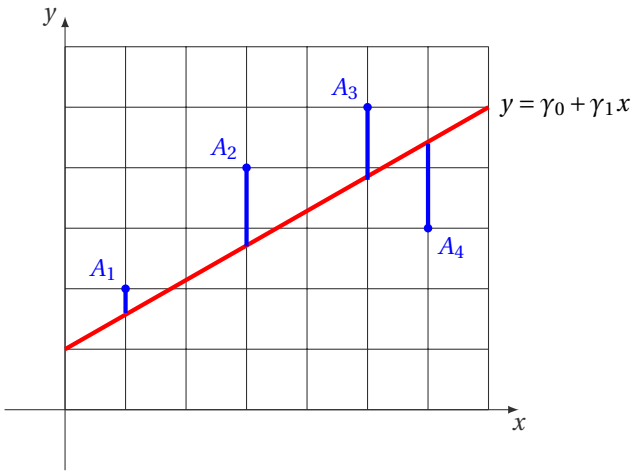
- ⋆ $r(\lambda\mathbf{x}, \lambda\mathbf{y}) = r(\mathbf{x},\mathbf{y})$ pour tout $\lambda \in \mathbb{R}^*$,
- ⋆ $r(\mathbf{x},\mathbf{y}) \in [-1;1]$.

8.5. Régression linéaire revisitée

L'ANALYSE DE RÉGRESSION donne des outils de prédiction du comportement d'un caractère si on connaît la valeur d'un autre caractère. L'ANALYSE DE CORRÉLATION mesure la force de la relation linéaire entre les deux caractères.
Considérons une série statistique bivariée $(x_k, y_k)_{k\in[\![1;n]\!]}$. On peut associer à chaque donnée (x_k, y_k) un point du plan et on peut représenter un échantillon de n données comme un nuage de n points. Si le nuage a une forme allongée, on peut essayer de dessiner une droite passant au milieu de ces points. Cette droite, appelée droite de régression linéaire, est un moyen de représenter la dépendance linéaire des deux caractères. La méthode des moindre carrés permet de déterminer la "meilleur" droite passante par le nuage de points constitué par une série statistique double.

8.5.1. Régression linéaire et moindres carrés

On considère un ensemble de N points $A_i = (x_i, y_i)$, $i = 1,\dots,N$. L'objectif est de trouver l'équation $y = \gamma_0 + \gamma_1 x$ de la droite qui approche au mieux tous ces points. Précisons ce que veut dire "approcher au mieux" : il s'agit de minimiser la somme des carrés des distances verticales entre les points et la droite.



La formule qui donne l'erreur est :

$$E(\gamma_0,\gamma_1) = \sum_{i=1}^{N} \big(y_i - (\gamma_0 + \gamma_1 x)\big)^2,$$

autrement dit

$$E(\gamma_0,\gamma_1) = \big(y_1 - (\gamma_0 + \gamma_1 x_1)\big)^2 + \cdots + \big(y_N - (\gamma_0 + \gamma_1 x_N)\big)^2.$$

Remarquons que l'on a toujours $E(\gamma_0,\gamma_1) \geq 0$. Si par exemple tous les points sont alignés, alors on peut trouver a et b tels que $E(\gamma_0,\gamma_1) = 0$. Quand ce n'est pas le cas, on cherche γ_0 et γ_1 qui rendent $E(\gamma_0,\gamma_1)$ le plus petit possible. Il s'agit donc bien ici de minimiser une fonction de deux variables (les variables sont γ_0 et γ_1). Pour cela nous aurons besoin de calculer son gradient :

$$\nabla E(\gamma_0,\gamma_1) = \left(\frac{\partial E}{\partial \gamma_0}(\gamma_0,\gamma_1), \frac{\partial E}{\partial \gamma_1}(\gamma_0,\gamma_1)\right) = \left(\sum_{i=1}^{N} -2\big(y_i - (\gamma_0 + \gamma_1 x_i)\big), \sum_{i=1}^{N} -2x_i\big(y_i - (\gamma_0 + \gamma_1 x_i)\big)\right).$$

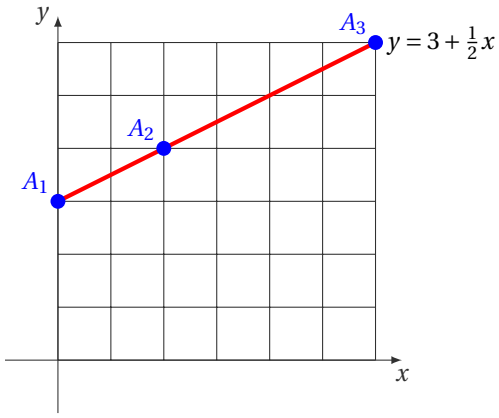
EXEMPLE
Prenons d'abord l'exemple des trois points $A_1 = (0,3)$, $A_2 = (2,4)$ et $A_3 = (6,6)$. La fonction $E(\gamma_0,\gamma_1)$ s'écrit :

$$E(\gamma_0,\gamma_1) = (3-\gamma_0)^2 + (4-(\gamma_0+2\gamma_1))^2 + (6-(\gamma_0+6\gamma_1))^2 = 40\gamma_1^2 + 16\gamma_1\gamma_0 - 88\gamma_1 + 3\gamma_0^2 - 26\gamma_0 + 61.$$

Ainsi

$$\nabla E(\gamma_0,\gamma_1) = \begin{pmatrix} 16\gamma_1 + 6\gamma_0 - 26 \\ 80\gamma_1 + 16\gamma_0 - 88 \end{pmatrix}$$

et $\nabla E(\gamma_0,\gamma_1) = \mathbf{0}$ ssi $\gamma_1 = \frac{1}{2}$ et $\gamma_0 = 3$. De plus, $E(3, \frac{1}{2}) = 0$ (les points sont alignés).



EXEMPLE
À partir des données des 5 points suivants, quelle ordonnée peut-on extrapoler pour le point d'abscisse $x = 6$?

$$A_1 = (4,1), \quad A_2 = (7,3), \quad A_3 = (8,3), \quad A_4 = (10,6), \quad A_5 = (12,7).$$

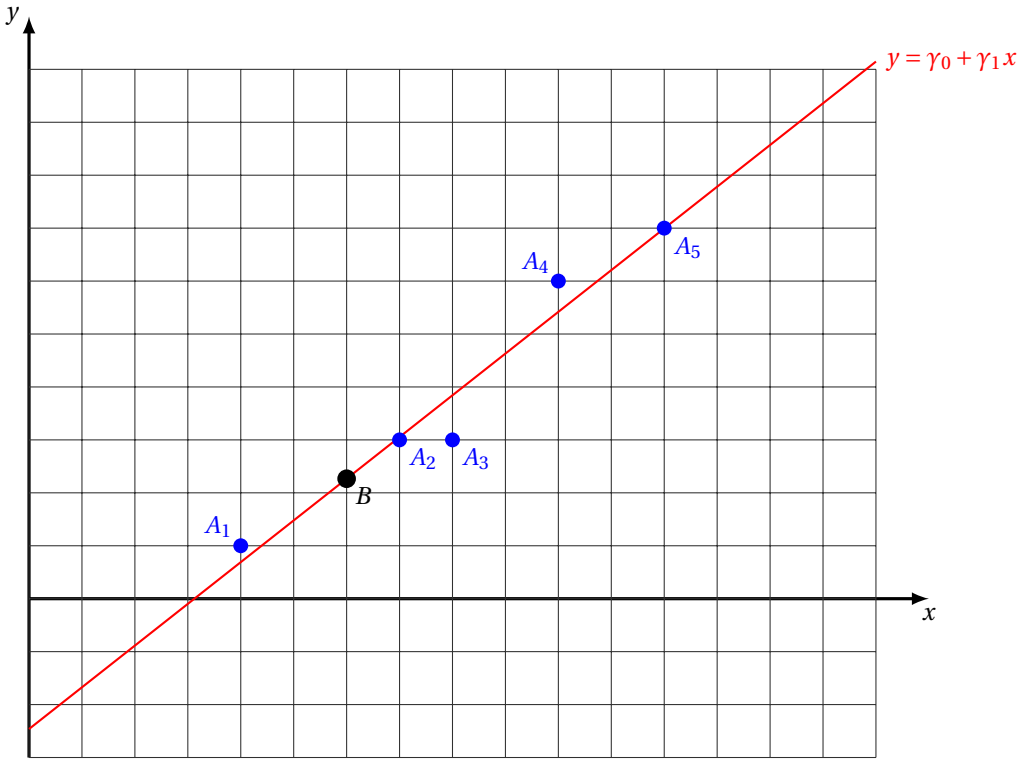
Ces 5 points sont à peu près alignés. On calcule la meilleure droite de régression linéaire en minimisant la fonction $E(\gamma_0, \gamma_1)$:

$$\begin{aligned} E(\gamma_0,\gamma_1) &= (-12\gamma_1 - \gamma_0 + 7)^2 + (-10\gamma_1 - \gamma_0 + 6)^2 + (-8\gamma_1 - \gamma_0 + 3)^2 + (-7\gamma_1 - \gamma_0 + 3)^2 + (-4\gamma_1 - \gamma_0 + 1)^2 \\ &= 373\gamma_1^2 + 82\gamma_1\gamma_0 - 386\gamma_1 + 5\gamma_0^2 - 40\gamma_0 + 104746\gamma_1 + 82\gamma_0 - 386. \end{aligned}$$

Ainsi

$$\nabla E(\gamma_0,\gamma_1) = \begin{pmatrix} 82\gamma_1 + 10\gamma_0 - 40 \\ 746\gamma_1 + 82\gamma_0 - 386 \end{pmatrix}$$

et $\nabla E(\gamma_0, \gamma_1) = \mathbf{0}$ ssi $\gamma_1 = \frac{145}{184} \approx 0.788$ et $\gamma_0 = -\frac{453}{184} \approx -2.462$. De plus, $E(-\frac{453}{184}, \frac{145}{184}) = 211/184 > 0$ (les points ne sont pas alignés).
Par conséquent, selon notre modèle linéaire, pour $x = 6$, on doit avoir $y = \gamma_0 + 6\gamma_1 = \frac{417}{184} \approx 2.27$ (le point B de la figure ci-dessus).



8.5.2. Droite de régression de y par rapport à x

On cherche à déterminer la droite d'équation $y = \gamma_0 + \gamma_1 x$ minimisant l'erreur quadratique $\mathscr{E} : \mathbb{R}^2 \to \mathbb{R}_+$ définie par

$$\mathscr{E}(\gamma_0, \gamma_1) = \sum_{k=1}^{n} \big(y_k - (\gamma_0 + \gamma_1 x_k)\big)^2$$

qui est la somme des distances au carré entre les points (x_k, y_k) et les points $(x_k, \gamma_0 + \gamma_1 x_k)$ de même abscisse situés sur la droite $y = \gamma_0 + \gamma_1 x$. Au chapitre 7 on a montré que γ_0 et γ_1 sont solution du système linéaire [1]

$$\begin{bmatrix} 1 & \frac{1}{n}\sum_{k=1}^n x_k \\ \frac{1}{n}\sum_{k=1}^n x_k & \frac{1}{n}\sum_{k=1}^n x_k^2 \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \gamma_1 \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\sum_{k=1}^n y_k \\ \frac{1}{n}\sum_{k=1}^n x_k y_k \end{bmatrix}$$

autrement dit, avec les notations introduites dans ce chapitre,

$$\begin{bmatrix} 1 & \bar{\mathbf{x}} \\ \bar{\mathbf{x}} & V(\mathbf{x}) + (\bar{\mathbf{x}})^2 \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \gamma_1 \end{bmatrix} = \begin{bmatrix} \bar{\mathbf{y}} \\ C(\mathbf{x}, \mathbf{y}) + \bar{\mathbf{x}}\bar{\mathbf{y}} \end{bmatrix}.$$

En résolvant ce système on trouve

$$\gamma_1 = \frac{C(\mathbf{x}, \mathbf{y})}{V(\mathbf{x})}, \qquad \text{coefficient directeur (pente)},$$
$$\gamma_0 = \bar{\mathbf{y}} - \gamma_1 \bar{\mathbf{x}}, \qquad \text{ordonnée à l'origine},$$

autrement dit $y = \gamma_1(x - \bar{\mathbf{x}}) + \bar{\mathbf{y}}$ (la droite passe par le point $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$).
D'un point de vue computationnel, cette écriture est susceptible de générer des erreurs de *roundoff* (les deux termes au numérateur ainsi qu'au dénominateur sont presque égaux, *i.e.* $C(\mathbf{x}, \mathbf{y})$ et $V(\mathbf{x})$ sont proches de zéro). Il est alors plus stable de calculer γ_1 comme suit (ce qui est équivalent) :

$$\gamma_1 = \frac{\sum_{k=0}^n \big(y_k(x_k - \bar{\mathbf{x}})\big)}{\sum_{k=0}^n \big(x_k(x_k - \bar{\mathbf{x}})\big)}.$$

8.5.3. Droite de régression de x par rapport à y

En échangeant les rôles de $\mathbf{x}$ et $\mathbf{y}$ on obtient la régression linéaire de $\mathbf{x}$ par rapport à $\mathbf{y}$. En générale les deux droites de régression sont distinctes.
En effet, dans le premier cas on minimise la somme des distances "verticales" (*i.e.* à x_i fixé), dans le deuxième cas il s'agit des distances "horizontale" (*i.e.* à y_i fixé) et en général ces deux quantités sont différentes.
Le produit des pentes de ces deux droites est égal à r^2 et les deux pentes sont égales si et seulement si $r = \pm 1$. Dans ce cas les deux droites coïncident et les points sont alignés.

8.5.4. Interprétation du coefficient de corrélation linéaire r

Il est toujours possible de tracer la droite des moindres carrés quelle que soit la forme du nuage. L'approximation du nuage par cette droite est-elle légitime ? Quel sens, quelle signification donner à cette droite ?
Dans un ajustement linéaire de $\mathbf{y}$ par rapport à $\mathbf{x}$ on appelle $\mathbf{x}$ la variable explicative (ou le "prédicteur") et $\mathbf{y}$ la variable expliquée (ou "à expliquer"). Le but d'un ajustement linéaire est d'expliquer une partie de la variation de $\mathbf{y}$ du fait de sa dépendance linéaire à $\mathbf{x}$.
Nous allons voir que le coefficient de corrélation r peut être utilisé pour mesurer la qualité d'une approximation de $\mathbf{y}$ par une fonction linéaire en x. Lorsque $r(\mathbf{x}, \mathbf{y})$ est en valeur absolue proche de 1 (en pratique strictement supérieur à 0.7), la droite de régression linéaire est une bonne approximation du nuage de point.

Notons $\hat{y}_k = \gamma_0 + \gamma_1 x_k$ pour $k = 1, \ldots, n$ la valeur estimée (ou prédite ou ajustée) de y_k par la régression linéaire lorsque $x = x_k$ et $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$. Il semble naturel de dire que remplacer le nuage par la droite trouvée est d'autant plus légitime que la dispersion du nuage de points par rapport à la droite des moindres carrés est petite. Autrement dit, on calcul l'erreur quadratique en son minimum (γ_0, γ_1) : l'approximation est légitime plus l'erreur quadratique $\mathscr{E}(\gamma_0, \gamma_1)$ est faible.

Soit γ_0 et γ_1 les valeurs qui minimisent l'erreur quadratique, alors

$$n\mathscr{E}(\gamma_0, \gamma_1) = \sum_{k=1}^{n} (y_k - \hat{y}_k)^2$$

1. NB : ici les indices commencent à 1 et on a tout divisé par n.

$$
\begin{aligned}
&= n\sum_{k=1}^{n}(y_k-\gamma_0-\gamma_1 x_k)^2\\
&= n\sum_{k=1}^{n}(y_k-(\bar{\mathbf{y}}-\gamma_1\bar{\mathbf{x}})-\gamma_1 x_k)^2\\
&= n\sum_{k=1}^{n}((y_k-\bar{\mathbf{y}})-\gamma_1(x_k-\bar{\mathbf{x}}))^2\\
&= n\sum_{k=1}^{n}(y_k-\bar{\mathbf{y}})^2+n\gamma_1^2\sum_{k=1}^{n}n(x_k-\bar{\mathbf{x}})^2-2n\gamma_1\sum_{k=1}^{n}(y_k-\bar{\mathbf{y}})(x_k-\bar{\mathbf{x}})\\
&= V(\mathbf{y})+\gamma_1^2 V(\mathbf{x})-2\gamma_1 C(\mathbf{x},\mathbf{y})\\
&= V(\mathbf{y})+\frac{C^2(\mathbf{x},\mathbf{y})}{V^2(\mathbf{x})}V(\mathbf{x})-2\frac{C(\mathbf{x},\mathbf{y})}{V(\mathbf{x})}C(\mathbf{x},\mathbf{y})\\
&= V(\mathbf{y})-\frac{C^2(\mathbf{x},\mathbf{y})}{V(\mathbf{x})}=V(\mathbf{y})\left(1-r^2(\mathbf{x},\mathbf{y})\right).
\end{aligned}
$$

Qualitativement, plus cette erreur est grande et moins bon est l'ajustement linéaire obtenu.
La quantité

$$SC_{\text{rés}}\overset{\text{def}}{=}\sum_{k=1}^{n}(y_k-\hat{y}_k)^2$$

est appelée **somme des carrés résiduelle** et est donc égale à

$$SC_{\text{rés}}=n\mathscr{E}(\gamma_0,\gamma_1)=V(\mathbf{y})\left(1-r^2(\mathbf{x},\mathbf{y})\right).$$

Elle est d'autant plus faible que r^2 est proche de 1. On peut alors interpréter l'erreur quadratique comme une mesure de la part de la variance de $\mathbf{y}$ qui ne peut pas être expliquée et prédite par une fonction linéaire en $\mathbf{x}$.
La variation totale

$$SC_{\text{tot}}\overset{\text{def}}{=}\sum_{k=1}^{n}(y_k-\bar{\mathbf{y}})^2$$

est appelée **somme des carrés totale** de $\mathbf{y}$ et est égale à

$$SC_{\text{tot}}=nV(\mathbf{y}).$$

On a donc

$$1-r^2(\mathbf{x},\mathbf{y})=\frac{SC_{\text{rés}}}{SC_{\text{tot}}},$$

i.e. la quantité $(1-r^2(\mathbf{x},\mathbf{y}))$ est égale à la proportion de variation de $\mathbf{y}$ non expliquée par la droite des moindres carrés.
La décomposition de la variation totale de $\mathbf{y}$ permet une autre interprétation de r^2 :

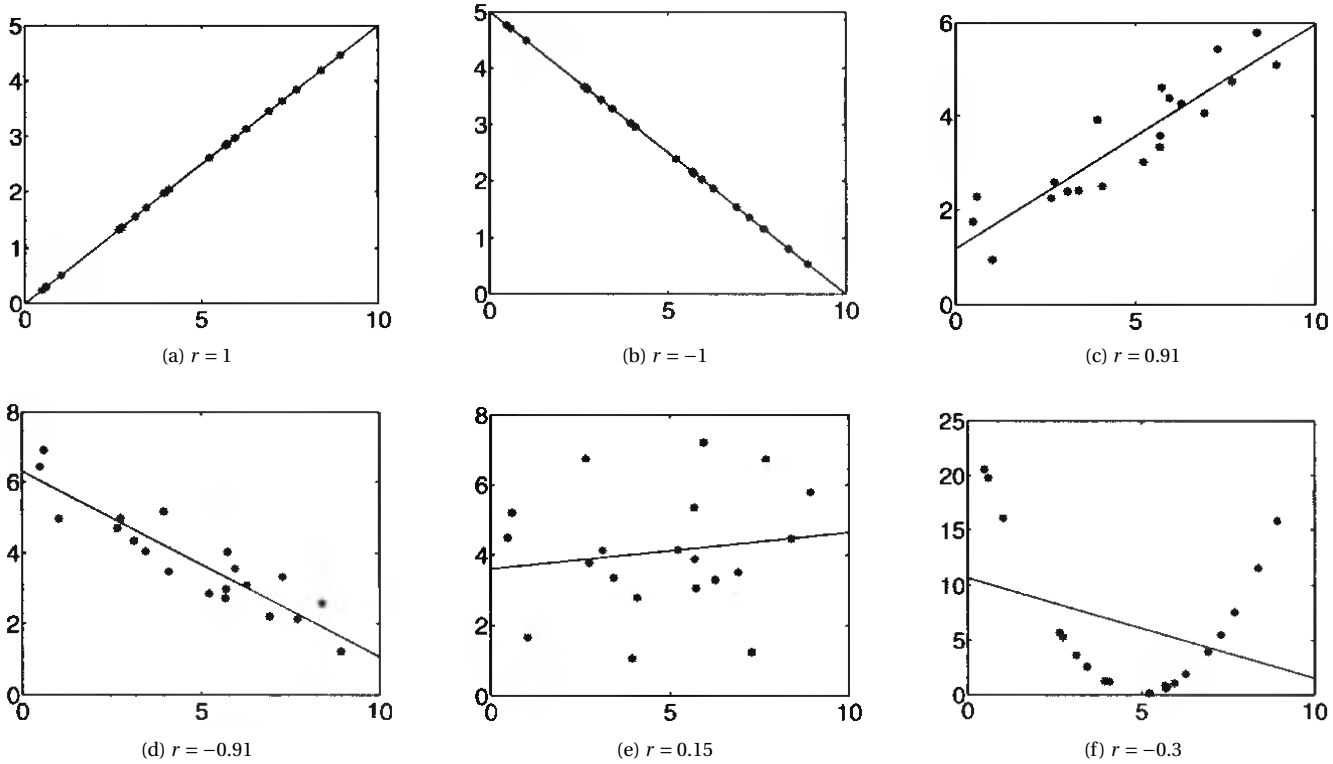
$$
\begin{aligned}
SC_{\text{tot}} &= \sum_{k=1}^{n}(y_k-\bar{\mathbf{y}})^2\\
&= \sum_{k=1}^{n}(y_k-\hat{y}_k+\hat{y}_k-\bar{\mathbf{y}})^2\\
&= \sum_{k=1}^{n}(y_k-\hat{y}_k)^2+\sum_{k=1}^{n}(\hat{y}_k-\bar{\mathbf{y}})^2+2\sum_{k=1}^{n}(y_k-\hat{y}_k)(\hat{y}_k-\bar{\mathbf{y}})\\
&= SC_{\text{rés}}+\sum_{k=1}^{n}(\hat{y}_k-\bar{\mathbf{y}})^2+2\sum_{k=1}^{n}(y_k-\hat{y}_k)(\hat{y}_k-\bar{\mathbf{y}}).
\end{aligned}
$$

Montrons que le dernier terme est nul :

$$
\begin{aligned}
\sum_{k=1}^{n}(y_k-\hat{y}_k)(\hat{y}_k-\bar{\mathbf{y}}) &= \gamma_1\left(\sum_{k=1}^{n}(y_k-\bar{\mathbf{y}})(x_k-\bar{\mathbf{x}})-\gamma_1\sum_{k=1}^{n}(x_k-\bar{\mathbf{x}})^2\right)\\
&= \gamma_1\left(C(\mathbf{x},\mathbf{y})-\gamma_1 V(\mathbf{x})\right)=0.
\end{aligned}
$$

On appelle **variation expliquée** par la régression la quantité

$$SC_{\text{expl}}=\sum_{k=1}^{n}(\hat{y}_k-\bar{\mathbf{y}})^2=nV(\hat{\mathbf{y}}).$$



(a) $r = 1$ (b) $r = -1$ (c) $r = 0.91$

(d) $r = -0.91$ (e) $r = 0.15$ (f) $r = -0.3$

FIGURE 8.3. – Qualité des ajustements linéaires en fonction du coefficient de corrélation. Ce coefficient reflète la non-linéarité et la direction d'une relation linéaire mais pas la pente de cette relation ni de nombreux aspects des relations non linéaires (dernière figure).

et on a l'égalité

$$SC_{\text{tot}} = SC_{\text{rés}} + SC_{\text{expl}}.$$

On a donc

$$r^2(\mathbf{x},\mathbf{y}) = \frac{SC_{\text{expl}}}{SC_{\text{tot}}},$$

i.e. $r^2(\mathbf{x},\mathbf{y})$ est égale à la proportion de variation de $\mathbf{y}$ expliquée par la droite des moindres carrés.

Le coefficient de corrélation r mesure la force et la direction de la relation entre $\mathbf{x}$ et $\mathbf{y}$. Deux cas extrêmes peuvent être facilement analysés :

- ⋆ si $r(\mathbf{x},\mathbf{y}) = \pm 1$, alors il existe un $\lambda_0 \in \mathbb{R}^*$ tel que $y_k - \bar{\mathbf{y}} = \lambda_0(x_k - \bar{\mathbf{x}})$ pour tout $k \in [\![1;n]\!]$. Cela montre que $\mathbf{x}$ et $\mathbf{y}$ sont parfaitement corrélés ;
- ⋆ si $r(\mathbf{x},\mathbf{y}) = 0$, alors la meilleur droite d'ajustement linéaire est la droite horizontale d'équation $y = \bar{\mathbf{y}}$ ce qui tend à montrer que les deux caractères ne sont pas corrélés.

La figure 8.3 donne plusieurs exemples pour différentes valeurs du coefficient de corrélation. Une valeur de $r(\mathbf{x},\mathbf{y})$ proche de 1 indique que les caractères sont positivement corrélés, et la meilleure droite d'ajustement linéaire obtenue par la méthode des moindres carrés a une pente positive. Une valeur de $r(\mathbf{x},\mathbf{y})$ proche de -1 indique que les caractères sont négativement corrélés, et la meilleure droite d'ajustement linéaire a une pente négative.

Noter que le coefficient de corrélation mesure seulement la qualité d'une relation linéaire : les caractères peuvent être corrélés mais pas linéairement, dans ce cas r sera petit et il faudrait généraliser ces notions aux cas des ajustements polynomiales.

Remarque (r : diviser par n ou $n-1$?)

Dans la définition de r, les dénominateurs utilisés pour la covariance $C(\mathbf{x},\mathbf{y})$ et pour les variances $V(\mathbf{x})$ et $V(\mathbf{y})$ sont n.

Si l'on tente d'estimer la corrélation de la population à partir d'un échantillon il faudra utiliser l'estimation de la covariance, toujours notée $C(\mathbf{x},\mathbf{y})$, ainsi que les estimations des variances $E(V(\mathbf{x}))$ et $E(V(\mathbf{y}))$, cela revient à diviser par $(n-1)$. Ce rapport

donne la même valeur que r :

$$\frac{E(C(\mathbf{x},\mathbf{y}))}{\sqrt{E(V(\mathbf{x}))E(V(\mathbf{y}))}} = \frac{\frac{n}{n-1}C(\mathbf{x},\mathbf{y})}{\sqrt{\frac{n}{n-1}V(\mathbf{x})\frac{n}{n-1}V(\mathbf{y})}} = \frac{C(\mathbf{x},\mathbf{y})}{\sqrt{V(\mathbf{x})V(\mathbf{y})}} = r(\mathbf{x},\mathbf{y}).$$

8.6. Corrélation et mises en garde

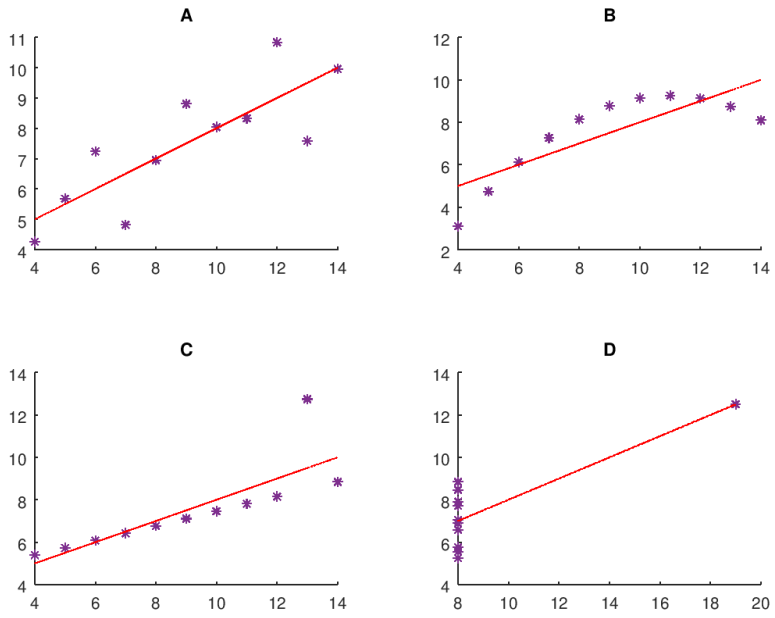
8.6.1. Le coefficient r et la qualité de l'ajustement linéaire

Comment juger la qualité de l'ajustement linéaire? Il est clair que si le coefficient r est voisin de 0, il faut rejeter l'ajustement linéaire, mais pour quelles valeurs de r, le considère-t-on de bonne qualité? C'est une question importante et beaucoup d'exemples montrent qu'on ne peut pas établir de règles de décision à partir du seul examen de la valeur de r.
Les exemples suivants montrent que le calcul du coefficient de corrélation linéaire doit toujours être complété par un examen graphique. Pour d'autres exemples voir par exemple https://www.autodesk.com/research/publications/same-stats-different-graphs

EXEMPLE
Considérons les quatre séries de 11 observations simultanées de deux variables **x** et **y** suivantes :

Série A		Série B		Série C		Série D	
x	**y**	**x**	**y**	**x**	**y**	**x**	**y**
10.00	8.04	10.00	9.14	10.00	7.46	8.00	6.58
8.00	6.95	8.00	8.14	8.00	6.77	8.00	5.76
13.00	7.58	13.00	8.74	13.00	12.74	8.00	7.71
9.00	8.81	9.00	8.77	9.00	7.11	8.00	8.84
11.00	8.33	11.00	9.26	11.00	7.81	8.00	8.47
14.00	9.96	14.00	8.10	14.00	8.84	8.00	7.04
6.00	7.24	6.00	6.13	6.00	6.08	8.00	5.25
4.00	4.26	4.00	3.10	4.00	5.39	8.00	5.56
12.00	10.84	12.00	9.13	12.00	8.15	19.00	12.50
7.00	4.82	7.00	7.26	7.00	6.42	8.00	7.91
5.00	5.68	5.00	4.74	5.00	5.73	8.00	6.89

On obtient grosso modo la même valeur du coefficient de corrélation linéaire ($r \approx 0.816$) et la même droite des moindres carrés $y \approx 3 + 0.5x$, mais l'examen graphique montre que l'ajustement linéaire n'est adapté qu'au premier cas.



EXEMPLE

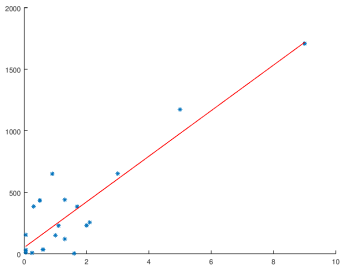
On se propose de calculer l'ajustement linéaire de la série de la composition minérale en fluorures et sodium (mg/l) de 21 eaux minérales gazeuses : [2]

Eau minérale	$\mathbf{x}$ = Fluorures	$\mathbf{y}$ = Sodium
Arcens	1.3	439
Arvie	0.9	650
Badoit	1	150
Beckerich	0.6	34
Châteauneuf	3	651
Eau de Perrier	0.05	11.5
Faustine	2	230
La Salvetat	0.25	7
Perrier	0.05	11.5
Puits St-Georges	0.5	434
Pyrénées	0.05	31
Quézac	2.1	255
San Pellegrino	0.6	35
St-Diéry	0.3	385
St-Jean	1.1	228
St-Pierre	1.7	383
St-Yorre	9	1708
Vernet	1.3	120
Vernière	0.05	154
Vichy-Célestins	5	1172
Wattwiller	1.6	3

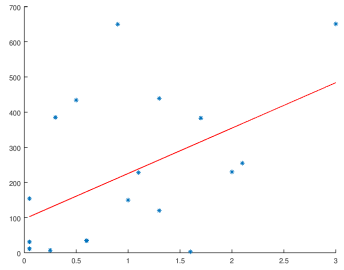
Calculons tout d'abord la moyenne et l'écart type :

$$\bar{\mathbf{x}} = 1.55, \qquad \bar{\mathbf{y}} = 338,$$
$$\sigma(\mathbf{x}) = 2.03, \qquad \sigma(\mathbf{x}) = 417.$$

Le coefficient de corrélation linéaire entre les deux composants minéraux est égal à 0.90. Cette valeur assez proche de 1 peut conduire à considérer que la droite des moindres carrés permet d'évaluer approximativement la teneur y en sodium en fonction de la teneur x en fluorures :



Cependant la représentation graphique du nuage des 21 points montre deux points caractérisés par une minéralité particulièrement élevée : «Vichy-Célestins» et «Saint-Yorre». Ces deux eaux minérales ont respectivement des valeurs « éloignée » et « extrême » pour les deux composants minéraux.En supprimant ces deux points et en réalisant l'ajustement sur les 19 autres points, on obtient :



2. Données extraites du journal "Que Choisir?", n° 422 bis, 2005

La moyenne et l'écart type sont maintenant

$$\bar{\mathbf{x}} = 0.97, \qquad \bar{\mathbf{y}} = 222,$$
$$\sigma(\mathbf{x}) = 0.81, \qquad \sigma(\mathbf{x}) = 208$$

et le coefficient r est passé de 0.9 à 0.5. Il faut aussi remarquer que les coefficients de la droite des moindres carrés sont passés respectivement de 185 à 129 et de 51 à 96.15.
Quel crédit apporter à un ajustement pour lequel deux points ont une telle influence ? On est donc obligé d'abandonner l'idée d'une relation linéaire entre les deux composants minéraux.

Tous ces résultats montrent qu'il ne faut jamais conclure sur la dépendance entre deux variables quantitatives au seul examen de la valeur du coefficient de corrélation linéaire.

De plus, lorsqu'une liaison linéaire entre deux variables a été mise en évidence par l'étude d'une série de n observations sur ce couple, il faut bien se garder de conclure à une relation de cause à effet entre ces variables sans en avoir examiné attentivement la signification, comme on va voir à la prochaine section.
L'examen graphique, ainsi que celui de la signification des variables, sont des compléments indispensables à l'information donnée par la valeur du coefficient de corrélation linéaire.

8.6.2. Distinguer causalité et corrélation

En statistiques, deux variables (choses que l'on mesure) sont corrélées positivement si elles évoluent de la même façon (augmentent en même temps, diminuent en même temps). Elles sont corrélées négativement si elles évoluent en sens inverse.
On établie un lien de causalité entre deux variables lorsqu'il y a un lien de cause à effet entre les deux, lorsque l'une est conséquence de l'autre.
L'**effet cigogne**[3] est une erreur qui consiste à confondre corrélation et causalité : «Deux variables évoluent de la même façon, l'une est donc forcément la cause de l'autre».

> «L'Alsace est la région de France où l'on observe le plus de cigognes. C'est également la région de France où il y a le plus de naissances. C'est donc la preuve que les cigognes apportent les bébés.»

Erreur si proche de l'effet cigogne qu'on les confond souvent, il s'agit ici de confondre succession et causalité[4] : «Deux évènements se suivent dans le temps, le premier est donc forcément la cause du second.»

EXEMPLE
Voici quelques exemples de ces deux confusions.

- Thomas met son caleçon rayé, puis il va au casino et gagne le gros lot. Il en conclue que son caleçon lui a porté chance.
- Plus les éoliennes tournent vite, plus y il a du vent : ce sont donc les éoliennes qui créent le vent !
- On constate que les pays où l'on mange le plus de viande sont les pays où l'on vit le plus longtemps. Doit-je changer mon régime alimentaire ? (On constate en réalité que ces pays sont également les plus riches, donc ceux où les habitants peuvent à la fois acheter plus de viande et avoir accès à de meilleurs soins)
- On constate que depuis que le parti de M. X est au pouvoir, le chômage diminue. Dois-je voter pour lui aux prochaines élections ? (Le chômage est lié à un grand nombre de facteurs très complexes, une simple corrélation est donc insuffisante pour démontrer que les actions de ce parti sont la cause de cette diminution. Il y a probablement un grand nombre de causes.)
- Je traînais un gros rhume depuis 3 jours, j'ai pris une tisane de camomille et le lendemain, j'allais mieux. La camomille m'a-t-elle guérie ? Ou bien est-ce j'aurais guéri de la même façon sans prendre de tisane, parce qu'un rhume se soigne généralement tout seul en 3 jours ?

Bien entendu, un corrélation peut donner des indices, interroger. Mais il ne s'agit en aucun cas d'un fait suffisant pour démontrer un lien de causes à effets. Pourtant, le raccourcit est rapide, instinctif, très largement utilisé dans les médias, et parfois très dangereux.
Une corrélation et une causalité sont deux objets distincts. Deux événements peuvent être corrélés sans pour autant avoir des rapports de cause à effet car d'autres variables pourraient être la cause des variations de **x** et de **y**.
Considérons par exemple l'affirmation suivante due à Coluche :

> «Quand on est malade, il ne faut surtout pas aller à l'hôpital : la probabilité de mourir dans un lit d'hôpital est 10 fois plus grande que dans son lit à la maison».

3. ou *Cum hoc, ergo propter hoc* : avec cela, donc à cause de cela.
4. ou *Post hoc, ergo propter hoc* : après cela, donc à cause de cela

Or, on ne meurt pas plus parce qu'on est dans un lit d'hôpital, mais on y est parce qu'on est malade, et quand on est malade la probabilité de mourir est plus grande.

Un autre exemple : une étude anglaise a prouvé que les gens habitant près de pylônes à haute tension étaient significativement plus souvent malades que le reste de la population. Est-ce la faute du courant électrique ? Ce n'est pas évident parce qu'une autre étude a révélé que les habitants sous les pylônes étaient en moyenne plus pauvres et on sait la corrélation (causalité ?) santé-pauvreté. À elle seule, cette étude ne permet pas de conclure.

Il en va ainsi des corrélations délinquance et origine ethnique : même à supposer qu'elles soient vraies, elles ne démontrent pas le rapport de cause à effet ; il peut se faire que la pauvreté, voire la détresse, soient liées à des discriminations ethniques, c'est alors cette misère qui est une cause possible de délinquance.

Démontrer une théorie avec seulement des statistiques peut être trompeur. Souvent la théorie préexiste et les chiffres sont ensuite utilisés pour la conforter «scientifiquement».

La corrélation relie les données et c'est ce que les big data brassent à très grosse échelle aujourd'hui. Ils accumulent une somme considérable de données et ils croisent tout ça en fonction de ce que l'on veut faire dire. Cependant, pour déterminer la nature du lien de causalité entre plusieurs éléments, c'est plus complexe. La théorie doit avoir un pouvoir explicatif, ne serait-ce que pour savoir dans quel sens lire les corrélations si jamais un lien de causalité existe. Il est par exemple maintenant bien établi qu'historiquement les variations de température sont étroitement liées aux variations de concentration de gaz carbonique dans l'atmosphère. Mais c'est la théorie qui permet de dire si c'est le réchauffement qui crée l'excès de gaz carbonique, ou l'inverse.

8.7. Exercices

Exercice 8.1 (Série univariée)
Une classe a été divisée en deux groupes de TP : le groupe TP_1 de $n_1 = 10$ étudiants et le groupe TP_2 de $n_2 = 4$ étudiants. Lors d'un contrôle noté sur 5, les étudiants du groupe TP_1 ont reçu les notes 4, 1, 3, 3, 4, 2, 3, 5, 3, 4 tandis que ceux du groupe TP_2 ont reçu les notes 4, 4, 4 et 5.
Pour chaque groupe κ, calculer la moyenne, le mode et la médiane des notes.
Calculer ensuite la moyenne, le mode et la médiane des notes de la classe.

Correction
Pour le groupe TP_1 on note $\mathbf{u} = (1,2,3,3,3,3,4,4,4,5)$ (dans l'ordre croissante) ce qui donne le tableau des fréquences

Note	Effectif (Nombre d'étudiants)	Fréquence (Proportion d'étudiants)
1	1	1/10
2	1	1/10
3	4	4/10
4	3	3/10
5	1	1/10
	$\sum = 10$	$\sum = 1$

Le mode est 3 (c'est la classe la plus importante). La moyenne vaut

$$\bar{\mathbf{u}} = \frac{1}{10}(1+2+3+3+3+3+4+4+4+5) = 3.2$$

soit encore, à partir du tableau,

$$\bar{\mathbf{u}} = \frac{1}{10}(1\times 1+2\times 1+3\times 4+4\times 3+5\times 1) = 3.2$$

Comme on a un nombre pair d'éléments (10), la médiane vaut

$$M(\mathbf{u}) = \frac{u_5+u_6}{2} = 3.$$

Pour le groupe TP_2 on note $\mathbf{v} = (4,4,4,5)$ (dans l'ordre croissante) ce qui donne le tableau des fréquences

Note	Effectif (Nombre d'étudiants)	Fréquence (Proportion d'étudiants)
1	0	0/10
2	0	0/10
3	0	0/10
4	3	3/10
5	1	1/10
	$\sum = 4$	$\sum = 1$

Le mode est 4 (c'est la classe la plus importante). La moyenne vaut

$$\bar{\mathbf{v}} = \frac{1}{4}(4+4+4+5) = 4.25$$

soit encore, à partir du tableau,

$$\bar{\mathbf{v}} = \frac{1}{4}(1\times 0 + 2\times 0 + 3\times 0 + 4\times 3 + 5\times 1) = 4.25$$

Comme on a un nombre pair d'éléments (4), la médiane vaut

$$M(\mathbf{v}) = \frac{v_2 + v_3}{2} = 4.$$

Pour la classe fusion des deux groupes de TP, on note $\mathbf{x} = (1,2,3,3,3,3,4,4,4,4,4,4,5,5)$ (dans l'ordre croissante) ce qui donne le tableau des fréquences

Note	Effectif (Nombre d'étudiants)	Fréquence (Proportion d'étudiants)
1	1	1/14
2	1	1/14
3	4	4/14
4	6	6/14
5	2	2/14
	$\sum = 14$	$\sum = 1$

Le mode est 4 (c'est la classe la plus importante). La moyenne vaut

$$\bar{\mathbf{x}} = \frac{1}{14}(1\times 1 + 2\times 1 + 3\times 4 + 4\times 6 + 5\times 2) = \frac{49}{14} = 3.5$$

soit encore, d'après la propriété sur la fusion de données,

$$\bar{\mathbf{x}} = \frac{n_1\bar{\mathbf{u}} + n_2\bar{\mathbf{v}}}{n_1+n_2} = \frac{10\times 3.2 + 4\times 4.25}{10+4} = \frac{49}{14} = 3.5$$

Comme on a un nombre pair d'éléments (14), la médiane vaut

$$M(\mathbf{x}) = \frac{x_7 + x_8}{2} = 4.$$

Exercice 8.2 (Covariance)

Calculer la covariance dans les cas suivants :

1. $\{(1,1),(-1,-1)\}$
2. $\{(-1,1),(1,-1)\}$
3. $\{(1,1),(-1,-1),(-1,1),(1,-1)\}$

Correction

1. On a $\mathbf{x} = (1,-1)$ et $\mathbf{y} = (1,-1)$ donc $\bar{\mathbf{x}} = \bar{\mathbf{y}} = 0$ et $\mathbb{C}(\mathbf{x},\mathbf{y}) = \frac{1\times 1 + (-1)\times(-1)}{2} - 0 = 1$: les points sont alignés et la pente est positive.
2. On a $\mathbf{x} = (-1,1)$ et $\mathbf{y} = (1,-1)$ donc $\bar{\mathbf{x}} = \bar{\mathbf{y}} = 0$ et $\mathbb{C}(\mathbf{x},\mathbf{y}) = \frac{(-1)\times 1 + 1\times(-1)}{2} - 0 = -1$: les points sont alignés et la pente est négative.

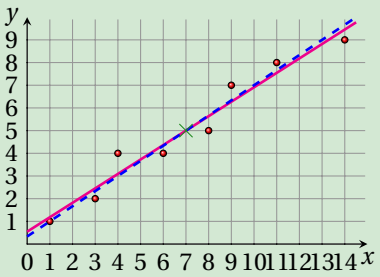
3. On a $\mathbf{x}=(1,-1,-1,1)$ et $\mathbf{y}=(1,-1,1,-1)$ donc $\bar{\mathbf{x}}=\bar{\mathbf{y}}=0$ et $\mathbb{C}(\mathbf{x},\mathbf{y})=\frac{1\times1+(-1)\times(-1)+(-1)\times1+1\times(-1)}{4}-0=0$: il n'y a pas de corrélation.

Exercice 8.3 (Régression linéaire)

Calculer les droites de meilleur approximation de l'ensemble de points suivant :

x	1	3	4	6	8	9	11	14
y	1	2	4	4	5	7	8	9

ainsi que leurs coefficients de corrélation.



Correction

Nous avons une série statistique double avec une population d'effectif $n=8$.
Pour calculer la droite de régression de $\mathbf{y}$ par rapport à $\mathbf{x}$ on calcule les quantités suivantes :

$$\bar{\mathbf{x}}=\frac{1}{n}\sum_{k=1}^{n}x_k=\frac{56}{8}=7$$
$$\bar{\mathbf{y}}=\frac{1}{n}\sum_{k=1}^{n}y_k=\frac{40}{8}=5$$
$$V(\mathbf{x})=\frac{1}{n}\sum_{k=1}^{n}x_k^2-\bar{\mathbf{x}}^2=\frac{33}{2}$$
$$C(\mathbf{x},\mathbf{y})=\frac{1}{n}\sum_{k=1}^{n}x_k y_k-\bar{\mathbf{x}}\bar{\mathbf{y}}=\frac{21}{2}$$
$$\gamma_1=\frac{C(\mathbf{x},\mathbf{y})}{V(\mathbf{x})}=\frac{7}{11}$$
$$\gamma_0=\bar{\mathbf{y}}-\gamma_1\bar{\mathbf{x}}=\frac{6}{11}$$
$$V(\mathbf{y})=\frac{1}{n}\sum_{k=1}^{n}y_k^2-\bar{\mathbf{y}}^2=7$$
$$r(\mathbf{x},\mathbf{y})=\frac{C(\mathbf{x},\mathbf{y})}{\sqrt{V(\mathbf{x})V(\mathbf{y})}}=\sqrt{\frac{21}{22}}>0.97$$

La droite cherchée a donc pour équation $y=\gamma_0+\gamma_1 x=\frac{6}{11}+\frac{7}{11}x$ avec une forte corrélation (mais cela ne dit rien sur la causalité entre les deux quantités!).

```
xx=[1,3,4,6,8,9,11,14]
yy=[1,2,4,4,5,7,8,9]

n = length(xx)
moy_x = mean(xx) %sum(xx)/n
moy_y = mean(yy) %sum(yy)/n
var_x = var(xx,1) %sum(xx.^2)/n-moy_x^2
var_y = var(yy,1) %sum(yy.^2)/n-moy_y^2
cov_xy = cov(xx,yy,1) %sum(xx.*yy)/n-moy_x*moy_y
gamma_1 = cov_xy/var_x
gamma_0 = moy_y-gamma_1*moy_x
r_xy = cov_xy / sqrt(var_x*var_y)
```

Pour calculer la droite de régression de $\mathbf{x}$ par rapport à $\mathbf{y}$ on calcule les quantités suivantes :

$$C(\mathbf{y},\mathbf{x})=C(\mathbf{x},\mathbf{y})=\frac{21}{2}$$
$$\gamma_1'=\frac{C(\mathbf{y},\mathbf{x})}{V(\mathbf{y})}=\frac{3}{2}$$
$$\gamma_0'=\bar{\mathbf{x}}-\gamma_1'\bar{\mathbf{y}}=-\frac{1}{2}$$

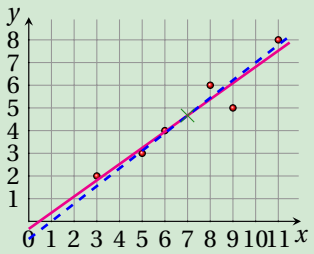
La droite cherchée a donc pour équation $x = \gamma_0' + \gamma_1' y = -\frac{1}{2} + \frac{3}{2} y$, soit encore $y = \frac{1}{3} + \frac{2}{3} x$.
On voit que $\gamma_1 \gamma_1' = \frac{7}{11}\frac{3}{2} = \frac{21}{22} = r^2$.

Exercice 8.4 (Régression linéaire)

Calculer les droites de meilleur approximation de l'ensemble de points suivant :

x	3	5	6	8	9	11
y	2	3	4	6	5	8

ainsi que leurs coefficients de corrélation.



Correction

Nous avons une série statistique double avec une population d'effectif $n = 6$.
Pour calculer la droite de régression de **y** par rapport à **x** on calcule les quantités suivantes :

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{k=1}^{n} x_k = \frac{42}{6} = 7$$
$$\bar{\mathbf{y}} = \frac{1}{n}\sum_{k=1}^{n} y_k = \frac{28}{6} = \frac{14}{3}$$
$$V(\mathbf{x}) = \frac{1}{n}\sum_{k=1}^{n} x_k^2 - \bar{\mathbf{x}}^2 = \frac{336}{6} - 49 = 7$$
$$C(\mathbf{x},\mathbf{y}) = \frac{1}{n}\sum_{k=1}^{n} x_k y_k - \bar{\mathbf{x}}\bar{\mathbf{y}} = \frac{226}{6} - 7\frac{14}{3} = 5$$
$$\gamma_1 = \frac{C(\mathbf{x},\mathbf{y})}{V(\mathbf{x})} = \frac{5}{7}$$
$$\gamma_0 = \bar{\mathbf{y}} - \gamma_1 \bar{\mathbf{x}} = -\frac{1}{3}$$
$$V(\mathbf{y}) = \frac{1}{n}\sum_{k=1}^{n} y_k^2 - \bar{\mathbf{y}}^2 = \frac{154}{6} - \frac{14^2}{9} = \frac{77\times 3 - 14^2}{9} = \frac{35}{9}$$
$$r(\mathbf{x},\mathbf{y}) = \frac{C(\mathbf{x},\mathbf{y})}{\sqrt{V(\mathbf{x})V(\mathbf{y})}} = \frac{3}{7}\sqrt{5} > 0.9$$

La droite cherchée a donc pour équation $y = \gamma_0 + \gamma_1 x = -\frac{1}{3} + \frac{5}{7} x$ avec une forte corrélation (mais cela ne dit rien sur la causalité entre les deux quantités!).

```
xx=[3,5,6,8,9,11]
yy=[2,3,4,6,5,8]

n = length(xx)
moy_x = mean(xx) %sum(xx)/n
moy_y = mean(yy) %sum(yy)/n
var_x = var(xx,1) %sum(xx.^2)/n-moy_x^2
var_y = var(yy,1) %sum(yy.^2)/n-moy_y^2
cov_xy = cov(xx,yy,1) %sum(xx.*yy)/n-moy_x*moy_y
gamma_1 = cov_xy/var_x
gamma_0 = moy_y-gamma_1*moy_x
r_xy = cov_xy / sqrt(var_x*var_y)
```

Pour calculer la droite de régression de **x** par rapport à **y** on calcule les quantités suivantes :

$$C(\mathbf{y},\mathbf{x}) = C(\mathbf{x},\mathbf{y}) = 5$$
$$\gamma_1' = \frac{C(\mathbf{y},\mathbf{x})}{V(\mathbf{y})} = \frac{9}{7}$$
$$\gamma_0' = \bar{\mathbf{x}} - \gamma_1' \bar{\mathbf{y}} = 1$$

La droite cherchée a donc pour équation $x = \gamma_0' + \gamma_1' y = 1 + \frac{9}{7} y$, soit encore $y = -\frac{7}{9} + \frac{7}{9} x$.
On voit que $\gamma_1 \gamma_1' = \frac{5}{7}\frac{9}{7} = \frac{45}{49} = r^2$.

Exercice 8.5

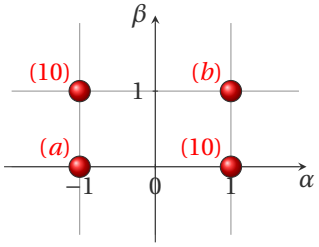
Soit le tableau de la distribution conjointe de deux variables quantitatives **x** et **y** :

$\mathscr{A}$ \ $\mathscr{B}$	$\beta_1 = 0$	$\beta_2 = 1$
$\alpha_1 = -1$ $\alpha_2 = 1$	$n_{1,1} = a$ $n_{2,1} = 10$	$n_{1,2} = 10$ $n_{2,2} = b$

1. Calculer les distributions marginales et écrire le tableau des fréquences de chaque couple et des fréquences marginales.
2. Calculer les distributions conditionnelles.
3. Calculer le coefficient de corrélation linéaire.

Correction

Ce tableau indique qu'on observe 10 fois le couple $(1,0)$, 10 fois le couple $(-1,1)$, a fois le couple $(-1,0)$ et b fois le couple $(1,1)$. On a donc au mieux $p \times q = 4$ points distincts (α_i, β_j) chacun avec un poids $n_{i,j}$:



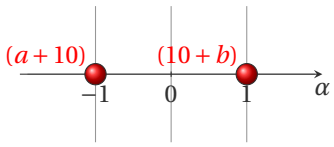
Si $a = b = 0$, alors on a seulement deux observations différentes sur deux variables (10 fois l'observation $(1,0)$ et 10 fois l'observation $(-1,1)$) : $r = -1$ (la droite de régression linéaire passe forcement par ces deux points et la pente est négative : la droite a pour équation $y = -\frac{1}{2}(x-1)$).
Si $a = b = 10$, il y a indépendance puisque les profils en lignes sont identiques donc $r = 0$ (la droite a pour équation $y = \frac{1}{2}$).
Si $a = 0$ et $b = 10$, il n'y a ni indépendance ($r \neq 0$), ni liaison linéaire ($r \neq \pm 1$). Même comportement si si $a = 10$ et $b = 0$.

Vérifions ce raisonnement par les calculs.

1. Distributions marginales
   - ⋆ Effectifs marginaux de α_i :

$$n_{1,\cdot} = 10 + a \qquad\qquad n_{2,\cdot} = 10 + b$$

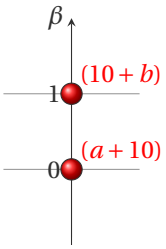
   et on a $\sum_{i=1}^{p=2} n_{i,\cdot} = 20 + a + b = n$. Autrement dit, indépendamment de l'observation de **y**, on observe $10 + a$ fois la valeur $x = \alpha_1 = -1$ et $10 + b$ fois la valeur $x = \alpha_2 = 1$.



   - ⋆ Effectifs marginaux de β_j :

$$n_{\cdot,1} = 10 + a \qquad\qquad n_{\cdot,2} = 10 + b$$

   et on a $\sum_{j=1}^{q=2} n_{\cdot,j} = 20 + a + b = n$. Autrement dit, indépendamment de l'observation de **x**, on observe $10 + a$ fois la valeur $y = \beta_1 = 0$ et $10 + b$ fois la valeur $y = \beta_2 = 1$.



⋆ Tableau des effectifs

$\mathscr{A}$ \ $\mathscr{B}$	$\beta_1 = 0$	$\beta_2 = 1$	Effectif marginal de α_i
$\alpha_1 = -1$	$n_{1,1} = a$	$n_{1,2} = 10$	$n_{1,\cdot} = 10 + a$
$\alpha_2 = 1$	$n_{2,1} = 10$	$n_{2,2} = b$	$n_{2,\cdot} = 10 + b$
Effectif marginal de β_j	$n_{\cdot,1} = 10 + a$	$n_{\cdot,2} = 10 + b$	$n = 20 + a + b$

Tableau des fréquences

$\mathscr{A}$ \ $\mathscr{B}$	$\beta_1 = 0$	$\beta_2 = 1$	Fréquence marginale de α_i
$\alpha_1 = -1$	$f_{1,1} = \frac{a}{20+a+b}$	$f_{1,2} = \frac{10}{20+a+b}$	$f_{1,\cdot} = \frac{10+a}{20+a+b}$
$\alpha_2 = 1$	$f_{2,1} = \frac{10}{20+a+b}$	$f_{2,2} = \frac{b}{20+a+b}$	$f_{2,\cdot} = \frac{10+b}{20+a+b}$
Fréquence marginale de β_j	$f_{\cdot,1} = \frac{10+a}{20+a+b}$	$f_{\cdot,2} = \frac{10+b}{20+a+b}$	1

2. Distributions conditionnelles :
   ⋆ De **x** sachant **y** :
      ⋆ De **x** sachant β_1 (on ne regarde que la colonne $y = \beta_1$) :

$$f_{i=1|j=1} = \frac{f_{i,j}}{f_{\cdot,j}} = \frac{a}{10+a}, \qquad \text{fréquence conditionnelle de } \alpha_1 \text{ sachant } \beta_1$$

$$f_{i=2|j=1} = \frac{f_{i,j}}{f_{\cdot,j}} = \frac{10}{10+a}, \qquad \text{fréquence conditionnelle de } \alpha_2 \text{ sachant } \beta_1$$

$$\bar{x}_{j=1} = f_{i=1|j=1}\alpha_1 + f_{i=2|j=1}\alpha_2 = \frac{10-a}{10+a}$$

      ⋆ De **x** sachant β_2 (on ne regarde que la colonne $y = \beta_2$) :

$$f_{i=1|j=2} = \frac{f_{i,j}}{f_{\cdot,j}} = \frac{10}{10+b}, \qquad \text{fréquence conditionnelle de } \alpha_1 \text{ sachant } \beta_2$$

$$f_{i=2|j=2} = \frac{f_{i,j}}{f_{\cdot,j}} = \frac{b}{10+b}, \qquad \text{fréquence conditionnelle de } \alpha_2 \text{ sachant } \beta_2$$

$$\bar{x}_{j=2} = f_{i=1|j=2}\alpha_1 + f_{i=2|j=2}\alpha_2 = \frac{10-b}{10+b}$$

      ⋆ Tableau des profiles en colonne $f_{i|j}$:

Profiles en colonne $f_{i\|j}$		
$\mathscr{A}$ \ $\mathscr{B}$	$\beta_1 = 0$	$\beta_2 = 1$
$\alpha_1 = -1$	$f_{1\|1} = \frac{a}{10+a}$	$f_{1\|2} = \frac{10}{10+b}$
$\alpha_2 = 1$	$f_{2\|1} = \frac{10}{10+a}$	$f_{2\|2} = \frac{b}{10+b}$
	1	1

   ⋆ De **y** sachant **x** :
      ⋆ De **y** sachant α_1 (on ne regarde que la ligne $x = \alpha_1$) :

$$f_{j=1|i=1} = \frac{f_{i,j}}{f_{i,\cdot}} = \frac{a}{10+a}, \qquad \text{fréquence conditionnelle de } \beta_1 \text{ sachant } \alpha_1$$

$$f_{j=2|i=1} = \frac{f_{i,j}}{f_{i,\cdot}} = \frac{10}{10+a}, \qquad \text{fréquence conditionnelle de } \beta_2 \text{ sachant } \alpha_1$$

$$\bar{y}_{i=1} = f_{j=1|i=1}\beta_1 + f_{j=2|i=1}\beta_2 = \frac{10}{10+a}$$

⋆ De $\mathbf{y}$ sachant α_2 (on ne regarde que la ligne $x = \alpha_2$) :

$$f_{j=1|i=2} = \frac{f_{i,j}}{f_{i,\cdot}} = \frac{10}{10+b}, \qquad \text{fréquence conditionnelle de } \beta_1 \text{ sachant } \alpha_2$$

$$f_{j=2|i=2} = \frac{f_{i,j}}{f_{i,\cdot}} = \frac{b}{10+b}, \qquad \text{fréquence conditionnelle de } \beta_2 \text{ sachant } \alpha_2$$

$$\bar{y}_{i=2} = f_{j=1|i=2}\beta_1 + f_{j=2|i=2}\beta_2 = \frac{b}{10+b}$$

⋆ Tableau des profiles en ligne $f_{j|i}$:

Profiles en ligne $f_{j\|i}$			
$\mathscr{A}$ \ $\mathscr{B}$	$\beta_1 = 0$	$\beta_2 = 1$	
$\alpha_1 = -1$	$f_{1\|1} = \frac{a}{10+a}$	$f_{1\|2} = \frac{10}{10+a}$	1
$\alpha_2 = 1$	$f_{2\|1} = \frac{10}{10+b}$	$f_{2\|2} = \frac{b}{10+b}$	1

3. Calcul du coefficient de corrélation linéaire r :

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{p} n_{i,\cdot}\alpha_i = \frac{(10+a)\times(-1)+(10+b)\times(1)}{20+a+b} = \frac{b-a}{20+a+b},$$

$$\bar{\mathbf{y}} = \frac{1}{n}\sum_{j=1}^{q} n_{\cdot,j}\beta_j = \frac{(10+a)\times(0)+(10+b)\times(1)}{20+a+b} = \frac{10+b}{20+a+b},$$

$$V(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{p} n_{i,\cdot}\alpha_i^2 - \bar{\mathbf{x}}^2 = \frac{(10+a)\times(-1)^2+(10+b)\times(1)^2}{20+a+b} - \frac{(b-a)^2}{(20+a+b)^2} = 1 - \frac{(b-a)^2}{(20+a+b)^2} = 4\frac{ab+10a+10b+100}{(20+a+b)^2},$$

$$V(\mathbf{y}) = \frac{1}{n}\sum_{j=1}^{q} n_{\cdot,j}\beta_i^2 - \bar{\mathbf{y}}^2 = \frac{(10+a)\times(0)^2+(10+b)\times(1)^2}{20+a+b} - \frac{(10+b)^2}{(20+a+b)^2} = \frac{10+b}{20+a+b}\left(1 - \frac{10+b}{20+a+b}\right) = \frac{(a+10)(b+10)}{(20+a+b)^2},$$

$$C(\mathbf{x},\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{i,j}\alpha_i\beta_j - \bar{\mathbf{x}}\bar{\mathbf{y}} = \frac{a\times(-1)\times(0)+10\times(-1)\times(1)+10\times(1)\times(0)+b\times(1)\times(1)}{20+a+b} - \frac{(b-a)(10+b)}{(20+a+b)^2}$$

$$= 2\frac{ab-100}{(20+a+b)^2},$$

$$r(\mathbf{x},\mathbf{y}) = \frac{C(\mathbf{x},\mathbf{y})}{\sqrt{V(\mathbf{x})V(\mathbf{y})}} = \frac{ab-100}{\sqrt{a+10}\sqrt{b+10}\sqrt{ab+10a+10b+100}}.$$

En particulier on voit que si $a = b = 0$ alors $r = -1$, si $a = 0$ et $b = 10$ alors $r = -\frac{1}{2}$, si $a = b = 10$ alors $r = 0$.
De plus, si on calcule les coefficients γ_1 de la régression linéaire de $\mathbf{y}$ en fonction de $\mathbf{x}$ (pente de la droite) on trouve :

$$\gamma_1 = \frac{C(\mathbf{x},\mathbf{y})}{V(\mathbf{x})} = \frac{1}{2}\frac{ab-100}{ab+10a+10b+100}$$

Si $a = b$ alors $\gamma_1 = \frac{a-10}{2(a+10)}$: en particulier, si $a = b = 10$ alors $\gamma_1 = 0$. Si $a = 0$ alors $\gamma_1 = \frac{-5}{b+10} < 0$: en particulier, si $b = 0$ alors $\gamma_1 = -\frac{1}{2}$. Même calcul si $b = 0$ car $\gamma_1 = \frac{-5}{a+10} < 0$.

☆ Exercice 8.6
Reproduire les graphes de l'exemple à la page 8.13.

Correction

```
xx_A=[10.0 8.0 13.0 9.0 11.0 14.0 6.0 4.0 12.0 7.0 5.0];
yy_A=[8.04 6.95 7.58 8.81 8.33 9.96 7.24 4.26 10.84 4.82 5.68];
```

```
xx_B=[10.0 8.0 13.0 9.0 11.0 14.0 6.0 4.0 12.0 7.0 5.0];
yy_B=[9.14 8.14 8.74 8.77 9.26 8.10 6.13 3.10 9.13 7.26 4.74];
xx_C=[10.0 8.0 13.0 9.0 11.0 14.0 6.0 4.0 12.0 7.0 5.0];
yy_C=[7.46 6.77 12.74 7.11 7.81 8.84 6.08 5.39 8.15 6.42 5.73];
xx_D=[8.0 8.0 8.0 8.0 8.0 8.0 8.0 8.0 19.0 8.0 8.0];
yy_D=[6.58 5.76 7.71 8.84 8.47 7.04 5.25 5.56 12.50 7.91 6.89];

T=[xx_A',yy_A',xx_B',yy_B',xx_C',yy_C',xx_D',yy_D'];
Cas=["A","B","C","D"];

for i=1:4
  CAS = Cas(i)
  xx = T(:,2*i-1);
  yy = T(:,2*i);
  r_xy = cov(xx,yy,1) / sqrt(var(xx,1)*var(yy,1))
  subplot(2,2,i)
  hold on
  plot(xx,yy,'*')
  gamma_1 = cov(xx,yy,1)/var(xx,1)
  gamma_0 = mean(yy)-gamma_1*mean(xx)
  d=@(x)gamma_0+gamma_1*x;
  plot(xx,d(xx),'r:')
  title(CAS)
  hold off
end
```

☆ Exercice 8.7
Reproduire les graphes de l'exemple à la page 8.14.

Correction

```
XX=[ 1.3 0.9 1 0.6 3 0.05 2 0.25 0.05 0.5 0.05 2.1 0.6 0.3 1.1 1.7 9 1.3 0.05 5 1.6];
YY=[439 650 150 34 651 11.5 230 7 11.5 434 31 255 35 385 228 383 1708 120 154 1172 3 ];

for i=1:2
  if i==1
    xx=XX;
    yy=YY;
  else % on enleve les deux valeurs extremes
    [val,idx]=max(YY);
    xx=XX([1:idx-1,idx+1:end]);
    yy=YY([1:idx-1,idx+1:end]);
    [val,idx]=max(yy);
    xx=xx([1:idx-1,idx+1:end]);
    yy=yy([1:idx-1,idx+1:end]);
  end
  figure()
  disp('')
  moy_x = mean(xx)
  moy_y = mean(yy)
  sigma_x = std(xx,1)
  sigma_y = std(yy,1)
  r_xy = cov(xx,yy,1) / sqrt(var(xx,1)*var(yy,1))
  hold on
  plot(xx,yy,'*')
  gamma_1 = cov(xx,yy,1)/var(xx,1)
  gamma_0 = mean(yy)-gamma_1*mean(xx)
  d=@(x)gamma_0+gamma_1*x;
  plot(xx,d(xx),'r:')
  hold off
end
```

ANNEXE A

Introduction à Octave/Matlab

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.**

A.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`). 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`), 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 *toolboxes* (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 toolboxes, 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`.

A.2. Installation(s) et version(s) en ligne

- La documentation et les sources d'Octave peuvent être téléchargées à l'adresse `https://www.gnu.org/software/octave/`.
  La version en ligne d'Octave est disponible ici `https://octave-online.net/`.
- L'université de Toulon propose aux étudiants la possibilité de le télécharger et de l'installer sur leur poste MATLAB. Toutes les informations sont ici `http://dsiun.univ-tln.fr/MATLAB.html`
  Par ailleurs, la version on line de MATLAB est disponible ici `https://fr.mathworks.com/products/matlab-online.html` Les étudiants et enseignants de l'université de Toulon peuvent s'y connecter avec leurs paramètres universitaires.

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».

A.3. 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".

A.3.1. Fenêtre de commandes : mode interactif

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 : [1]

```
>> A = [2 1 0; -1 2 2; 0 1 4];
>> b = [1; 2; 3];
>> soln = A\b
soln =
   0.25000
   0.50000
   0.62500
```

Ce mode interactif est très pratique pour rapidement tester des instructions et directement voir leurs résultats. Son utilisation reste néanmoins limitée à des programmes de quelques instructions. En effet, devoir à chaque fois retaper toutes les instructions s'avérera vite pénible.
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 préalablement écrit un fichier avec uniquement les commandes qu'on a tapées et l'avoir enregistré sur l'ordinateur avec l'extension `.m`. Une fois cela fait, on demandera à Octave de lire ce fichier et exécuter son contenu, instruction par instruction, comme si on les avait tapées l'une après l'autre dans la Fenêtre de commandes. Ainsi plus tard on pourra ouvrir ce fichier et lancer Octave sans avoir à retaper toutes les commandes. Passons alors à l'onglet "Éditeur".

A.3.2. Éditeur : mode script

On voit qu'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 dans un fichier avec l'**extension** `.m`, pour le faire tourner et afficher les résultats dans la "Fenêtre de commandes" il suffira d'appuyer sur la touche «F5». Si on a fait une faute de frappe, Octave le remarquera et demandera de corriger.
Maintenant qu'on a sauvé le programme, on est capable de le recharger.
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. Copier par exemple les lignes suivantes dans un fichier appelé `first.m`

```
A = [2 1 0; -1 2 2; 0 1 4];
b = [1; 2; 3];
soln = A\b
```

Appuyer sur la touche «F5», cliquer sur "Changer de répertoire" et regarder le résultat dans l'onglet "Fenêtre de commandes". [2]

A.4. Notions de base

A.4.1. Variables et affectation

Une variable peut être vue comme une boîte représentant un emplacement en mémoire qui permet de stocker une valeur et à qui on a donné un nom afin de facilement l'identifier (boîte ← valeur) :

```
>> x=1
x = 1
```

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

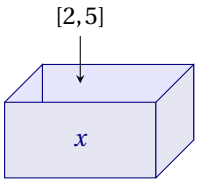
```
>> x='c'
x = c
```

1. Ces instructions calculent la solution du système linéaire $\begin{pmatrix} 2 & 1 & 0 \\ -1 & 2 & 2 \\ 0 & 1 & 4 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$. 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.

2. Sinon, si ce fichier se trouve dans le répertoire courant d'Octave, pour l'exécuter on peut juste taper son nom **(sans l'extension)** sur la ligne de commande d'Octave : `>> first`

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/first.m")`

L'affectation `x=[2 5]` crée une association entre le nom `x` et le vecteur $[2,5]$: la boîte de nom `x` contient le vecteur $[2,5]$.



Il faut bien prendre garde au fait que **l'instruction d'affectation** (=) **n'a pas la même signification que le symbole d'égalité (=) en mathématiques** (ceci explique pourquoi l'affectation de `1` à `x`, qu'en Octave s'écrit `x = 1`, en algorithmique se note souvent $x \leftarrow 1$).
Une fois une variable initialisée, on peut modifier sa valeur en utilisant de nouveau l'opérateur d'affectation (=). La valeur actuelle de la variable est remplacée par la nouvelle valeur qu'on lui affecte. Dans l'exemple précédent, on initialise une variable à la valeur 1 et on remplace ensuite sa valeur par le vecteur $[1,2]$.
Il est très important de donner un nom clair et précis aux variables. Par exemple, avec des noms bien choisis, on comprend tout de suite ce que calcule le code suivant :

```
base = 8
hauteur = 3
aire = base * hauteur / 2
```

Octave distingue les majuscules des minuscules. Ainsi `mavariable`, `Mavariable` et `MAVARIABLE` sont des variables différentes.
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). Cependant, certains mots sont réservés :

`ans`	Nom pour les résultats
`eps`	Le plus petit nombre tel que 1+eps>1
`inf`	∞
`NaN`	Not a number
`i` ou `j`	i
`pi`	π

```
>> 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 NaN's are not equal!
ans = 0
>> eps
ans = 2.2204e-16
```

Si on écrit une instruction sans affectation, le résultat sera affecté à la variable `ans`.

```
>> [4,3]
ans =

   4 3
>> 'Ciao'
ans = Ciao
```

Pour effacer la mémoire et désaffecter toutes les variables, utiliser la fonction `clear all`.

A.5. Commentaires

Le symbole `%` indique le début d'un **commentaire** : tous les caractères entre `%` et la fin de la ligne sont ignorés par l'interpréteur.
Dans l'éditeur d'Octave, pour commenter plusieurs lignes en même temps, les sélectionner et appuyer sur les touches «Ctrl+R». Pour dé-commenter plusieurs lignes en même temps, les sélectionner et appuyer sur les touches «Ctrl+Maj+R».

A.6. Affichage

Lors de l'affectation d'une variable, le résultat de l'affectation sera affiché ; le symbole `;` supprime cet affichage.

```
>> a=[1,2]
a =

   1 2

>> a=[4,3];
>> 'Ciao'
ans = Ciao
```

Pour afficher seulement le **contenu** d'une variable utiliser la fonction `disp` (en effet, si on écrit juste le nom de la variable, on affichera aussi le nom de la variable)

```
>> a=[4,3];
>> disp(a)
   4 3
>> a
a =

   4 3
>> disp('Ciao')
Ciao
```

Pour nettoyer la fenêtre de commandes, utiliser la fonction `clc`.

A.7. Opérations arithmétiques

Dans Octave on a les opérations arithmétiques usuelles :

`+`	Addition
`-`	Soustraction
`*`	Multiplication
`/`	Division
`^`	Exponentiation

Quelques exemples :

```
>> a = 100
a = 100
>> b = 17
b = 17
```

```
>> c = a-b
c = 83
>> a/b
ans = 5.8824
```

```
>> a^b
ans = 1.0000e+34
```

Les opérateurs arithmétiques possèdent chacun une priorité qui définit dans quel ordre les opérations sont effectuées. Par exemple, lorsqu'on écrit `1 + 2 * 3`, la multiplication va se faire avant l'addition. Le calcul qui sera effectué est donc `1 + (2 * 3)`. Dans l'ordre, l'opérateur d'exponentiation est le premier exécuté, viennent ensuite les opérateurs *, /, // et %, et enfin les opérateurs + et -.
Lorsqu'une expression contient plusieurs opérations de même priorité, ils sont évalués de gauche à droite. Ainsi, lorsqu'on écrit `1 - 2 - 3`, le calcul qui sera effectué est `(1 - 2) - 3`. En cas de doutes, vous pouvez toujours utiliser des parenthèses pour rendre explicite l'ordre d'évaluation de vos expressions arithmétiques.
Il existe aussi les opérateurs augmentés :

`a += b`	équivaut à a = a+b
`a -= b`	équivaut à a = a-b
`a *= b`	équivaut à a = a*b
`a /= b`	équivaut à a = a/b
`a ^ = b`	équivaut à a = a^ b

A.8. Division euclidienne

Lorsqu'on divise un nombre entier D (appelé dividende) par un autre nombre entier d (appelé diviseur), on obtient deux résultats : un quotient q et un reste r, tels que $D = qd + r$ (avec $r < d$). La valeur q est le résultat de la division entière et

la valeur r celui du reste de cette division. Par exemple, si on divise 17 par 5, on obtient un quotient de 3 et un reste de 2 puisque $17 = 3 \times 5 + 2$. Ces deux opérateurs sont très utilisés dans plusieurs situations précises. Par exemple, pour déterminer si un nombre entier est pair ou impair, il suffit de regarder le reste de la division entière par deux. Le nombre est pair s'il est nul et est impair s'il vaut 1. Une autre situation où ces opérateurs sont utiles concerne les calculs de temps. Si on a un nombre de secondes et qu'on souhaite le décomposer en minutes et secondes, il suffit de faire la division par 60. Le quotient sera le nombre de minutes et le reste le nombre de secondes restant. Par exemple, 175 secondes correspond à `175//60=2` minutes et `175%60=55` secondes.

```
>> q=fix(9/4)
q = 2
>> % Reste de la division euclidienne de 9 par 4
>> r=rem(9,4)
r = 1
>>
>> r=mod(9,4) % 9 modulo 4
r = 1
>> q=fix(175/60)
q = 2
>> r=rem(175,60)
r = 55
```

A.9. 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). Notons que le symbole `;` a deux fonctions : il supprime l'affichage d'un résultat intermédiaire et il sépare les lignes d'une matrice. 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 2×3 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
```

En Octave, les éléments d'une matrice sont *indexés à partir de* 1. 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. On peut extraire une sous-matrice en déclarant l'indice de **début (inclus)** et l'indice de **fin (inclus)**, séparés par deux-points `A(i:j)`, ou encore une sous-matrice en déclarant l'indice de début (inclus), l'indice de fin (inclus) et le pas, séparés par des deux-points `A(i:j:k)`. On peut même utiliser un pas négatif. Cette opération est connue sous le nom de *slicing* (en anglais).

```
A(2,3)   % element A_{23}
A(:,3)   % vecteur colonne [A_{13};...;A_{n3}]
A(1:4,3)  % [A_{13};...A_{43}] premieres 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 a la ligne 2 colonne 3
ans = 7
>> A(:,2) % Toutes les lignes, deuxieme colonne
ans =
   1
   5
   9

>> A(2:3,2:3) % Sous-matrice 2 x 2
ans =
   5 7
   9 2

>> A(3:-1:1,:) % les lignes de la derniere a la premiere, toutes les colonnes
ans =
   4 9 2
   3 5 7
   8 1 6
```

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

A.9.1. Matrices particulières

Construction de matrices particulières :

① 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)`.

② 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)`.

③ 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 $\mathbb{I}$.

④ La commande `A=[]` définit une matrice vide.

⑤ La commande `diag(v)` où $\mathbf{v}$ est un vecteur de n éléments renvoie une matrice carré de taille n dont les éléments valent 0 exceptés ceux de la diagonale principale qui valent $\mathbf{v}$.

⑥ Soit $\mathbf{v}$ un vecteur de n composantes. La commande `diag(v)` renvoie une matrice diagonale carrée de dimension n qui contient $\mathbf{v}$ sur la diagonale principale ; la commande `diag(v,1)` renvoie une matrice carrée de dimension $n+1$ qui contient $\mathbf{v}$ sur la sur-diagonale principale etc.

Notons que `diag(ones(1,4))` équivaut à `eye(4)`.

```
>> Z=zeros(2,3)
Z =
   0 0 0
   0 0 0
```

```
>> O=ones(3,2)
O =
   1 1
   1 1
   1 1
```

```
>> E=eye(2,5)
E =
Diagonal Matrix
   1 0 0 0 0
   0 1 0 0 0
```

```
>> A=[]
A = [](0x0)

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

```
>> F=diag(v)
F =
Diagonal Matrix
   1 0 0
   0 2 0
   0 0 3
```

```
>> G=diag(v,1)
G =
   0 1 0 0
   0 0 2 0
   0 0 0 3
   0 0 0 0
```

Construction de vecteurs :

① `x=[debut:pas:fin]`

② `x=linspace(debut,fin,N)` (**x** a N points donc le pas h est $\frac{x_{\text{fin}}-x_{\text{debut}}}{N-1} = \frac{x_N-x_1}{N-1}$)

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

Notons que la première instruction ne garantit pas que le dernier point soit pris, cela dépend du pas choisi (et des erreurs d'arrondis) :

```
x = [0 : 0.4 : 1] % output : x=0.00000 0.40000 0.80000
```

Dans la première instructions on peut utiliser un pas négatif :

```
x = [1 : -0.4 : 0] % output : x=1.00000 0.60000 0.20000
```

Dimensions :

```
A=eye(3,4);
[r,c]=size(A) % r=nb de lignes et c=nb de colonnes de A
x=[0:10];
n=length(x) % n=nb d'elements de x
```

A.9.2. Opérations entre matrices

Opérations sur les matrices (lorsque les dimensions sont compatibles) :

- ⋆ Somme $\mathbb{C} = \mathbb{A} + \mathbb{B}$, *i.e.* $\mathbb{C}_{ij} = \mathbb{A}_{ij} + \mathbb{B}_{ij}$: `C=A+B`
- ⋆ Produit $\mathbb{C} = \mathbb{A}\mathbb{B}$, *i.e.* $\mathbb{C}_{ij} = \sum_{k=1} \mathbb{A}_{ik} + \mathbb{B}_{kj}$: `C=A*B` NB il s'agit du **produit matriciel**!
- ⋆ Division à droite $\mathbb{C} = \mathbb{A}\mathbb{B}^{-1}$: `C=A/B`
- ⋆ Division à gauche $\mathbb{C} = \mathbb{A}^{-1}\mathbb{B}$: `C=A\B` (si $\mathbb{B}$ est un vecteur colonne alors $\mathbb{C}$ est un vecteur colonne **solution du système linéaire** $\mathbb{A}\mathbb{C} = \mathbb{B}$)
- ⋆ Élévation à la puissance $\mathbb{C} = \mathbb{A}\mathbb{A}\mathbb{A}$: `C=A^3`
- ⋆ Calcul du déterminant (si la matrice est carrée) : `det(A)`
- ⋆ Calcul de la matrice inverse (si la matrice est inversible) : `inv(A)`

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

```
>> B=ones(2,3)
B =
   1 1 1
   1 1 1
```

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

```
>> D=eye(3,2)
D =
Diagonal Matrix
   1 0
   0 1
   0 0
```

```
>> E=A(1:2,1:2)
E =
   1 2
   4 5
```

```
>> A+B
ans =
   2 3 4
   5 6 7

>> A*C
ans =
   22 28
   49 64

>> A/B
```

```
ans =
   1.00000 1.00000
   2.50000 2.50000

>> b=[28;64]
b =
   28
   64

>> A\b
ans =
   2.0000
   4.0000
   6.0000

>> A\B
ans =
  -5.0000e-01 -5.0000e-01 -5.0000e-01
   8.3267e-17 8.3267e-17 8.3267e-17
   5.0000e-01 5.0000e-01 5.0000e-01

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

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

```
>> A+C
error: operator +: nonconformant arguments (op1 is 2x3, op2 is 3x2)
```

A.9.3. 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 la multiplication élément par élément `.*` **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 $\mathbb{C}_{ij} = \mathbb{A}_{ij}\mathbb{B}_{ij}$: `C=A.*B` NB il s'agit du produit d'Hadamard et non pas du produit matriciel
- ⋆ Division $\mathbb{C}_{ij} = \mathbb{A}_{ij}/\mathbb{B}_{ij}$: `C=A./B`
- ⋆ Élévation à la puissance $\mathbb{C}_{ij} = \mathbb{A}^3_{ij}$: `C=A.^3`

Ces opérations sont à la base de la "vectorisation" car elles permettent le remplacement d'un boucle par une opération matricielle pointée qui est généralement beaucoup plus performante.

```
>> A=[1 2 3; 4 5 6]
A =
   1 2 3
   4 5 6
>> B=[0 2 0; 0 0 1]
B =
   0 2 0
   0 0 1
>> A.*B
ans =
   0 4 0
   0 0 6
>> A*B
error: operator *: nonconformant arguments (op1 is 2x3, op2 is 2x3)
```

A.9.4. Opérateurs de comparaison et connecteurs logiques

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

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

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

ATTENTION

Les opérateurs de comparaison agissent élément par élément, ainsi lorsqu'on les applique à une matrice le résultat est une matrice qui contient que des 0 ou 1 (parfois appelée "**masque**"). 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
```

On peut utiliser un masque pour remplacer seulement les éléments qui satisfont une conditions :

```
>> A = [1 -2 3; 4 -5 6];
>> A(A<0)=-100
A =
     1 -100 3
     4 -100 6
```

Les masques sont à la base de la "vectorisation" car permettent le remplacement d'un boucle avec des conditions par une opération matricielle qui est généralement beaucoup plus performante.

EXEMPLE (MASQUE)

Étant donné trois vecteurs :

- `hotels` : une liste de noms d'hôtels
- `ratings` : leurs notes dans une ville
- `cutoff` : la note minimale

on souhaite afficher les noms des hôtels ayant une note supérieure ou égale au seuil.

```
>> hotels =["CityLights";"SeaView";"MarketPlace";"ResortSpa";"Nightingale";"Clubadub";"SkylineView";"
    MarinaBay";"ComfortFirst";"VillageValley"]; % vecteur colonne
>> ratings = [7.2;8.7;6.5;9.3;4.3;6.9;8.8;5.9;7.4;9.1]; % vecteur colonne
>> cutoff = 8;
>> good = hotels(ratings>=cutoff,:) % NB ":" pour selectionner toute la chaine de caracteres
good =

SeaView
ResortSpa
SkylineView
VillageValley
```

Pour combiner des conditions complexes (par exemple $x > -2$ et $x^2 < 5$), on peut combiner les opérateurs de comparaison avec les connecteurs logiques :

On écrit	Ça signifie
&	et
|	ou
~	non

Par exemple

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

A.10. Fonctions

A.10.1. Fonctions prédéfinies

De très nombreuses fonctions sont déjà disponibles dans Octave/Matlab. Voici quelques exemples de fonction mathématique :

```
abs(-5)

sin(pi)
cos(pi)
tan(pi)

factorial(5) % output : 120

r=rem(11,3)

round(3.7) % output : 4
round(3.3) % output : 3
round(-3.7) % output : -4
round(-3.3) % output : -3

fix(3.7) % output : 3
fix(3.3) % output : 3
fix(-3.7) % output : -3
fix(-3.3) % output : -3

floor(3.7) % output : 3
floor(3.3) % output : 3
floor(-3.7) % output : -4
floor(-3.3) % output : -4

ceil(3.7) % output : 4
ceil(3.3) % output : 4
ceil(-3.7) % output : -3
ceil(-3.3) % output : -3
```

En générale les fonctions prédéfinies sont vectorisées, autrement dit si on applique la fonction à une matrice, elle renvoie une matrice de la même taille en ayant appliqué la fonction à chaque élément.

```
x=[1:5]

sqrt(x) % output 1.0000 1.4142 1.7321 2.0000 2.2361
exp(x) % output 2.7183 7.3891 20.0855 54.5982 148.4132
log(x) % output 0.00000 0.69315 1.09861 1.38629 1.60944
log10(x) % output 0.00000 0.30103 0.47712 0.60206 0.69897
log2(x) % output 0.00000 1.00000 1.58496 2.00000 2.32193

ismember(2,x) % output 1
ismember(11,x) % output 0
```

Certaines fonctions sont spécifiques aux matrices :

```
A=[1 2; 3 4];
det(A)
inv(A)
trace(A)
size(A) % dimensions de A
x=[1 2 3];
length(x) % longueur d'un vecteur
numel(x) % nombre d'elements d'un vecteur
```

A.10.2. Définition d'une fonction

On peut définir nos propres fonctions au moyen de la commande `function` et les utiliser dans plusieurs scripts.

```
function [y1,...,yN] = myfunc(x1,...,xM)
   instruction_1
   instruction_2
   ...
   [y1,...,yN] = ...
end
```

La structure type d'une fonction est la suivante :

- ⋆ on utilise la commande `function` dans laquelle on indique les arguments `(x1,...,xM)` et la valeur de retour `[y1,...,yN]`
- ⋆ cette déclaration est suivie du corps de la définition qui est un bloc d'instructions à exécuter et se termine par le mot-clé `end` ou `endfunction`

ATTENTION
Pour éviter de surcharger une fonction déjà définie dans Matlab/Octave, prendre l'habitude d'appeler ses fonctions par `my...`

Voir aussi https://fr.mathworks.com/help/matlab/matlab_prog/create-functions-in-files.html

Fichiers fonctions

Par convention, **chaque définition de fonction est stockée dans un fichier séparé qui porte le nom de la fonction** suivi de l'**extension** `.m` Ces fichiers s'appellent des fichiers de fonction. Notez que c'est la même extension que les fichiers de scripts mais, de plus, **il faut absolument que le fichier s'appelle comme la fonction qu'il contient**.
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 myfunc`
- ⋆ la fonction elle-même.

ATTENTION
Pour éviter toute confusion, utilisez le même nom pour le fichier de fonction et la première fonction du fichier. Matlab/Octave associe votre programme au nom du fichier, pas au nom de la fonction. Les fichiers de script ne peuvent pas avoir le même nom qu'une fonction du fichier.

Remarque (Sous-fonctions)
Un fichier de fonction peut en réalité contenir plusieurs fonctions 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.

Voici un exemple qui prend en entrée un vecteur de valeurs et renvoie la moyenne et la déviation standard :

Fichier `stat.m`

```
% Cette fonction calcule la moyenne et la deviation
% standard d'un vecteur x
function [m,s] = stat(x)
   n = length(x);
   m = sum(x)/n;
   s = sqrt(sum((x-m).^2/n));
end
```

Script

```
values = [12.7, 45.4, 98.9, 26.6, 53.1];
[average,stdeviation] = stat(values)
```

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 l'aire s d'un triangle par la formule de Heron.
% a, b, c sont les longueurs des aretes.
function s = heron(a, b, c)
   p = (a+b+c)/2;
   s = sqrt(p*(p-a)*(p-b)*(p-c));
endfunction
```

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
```

Fonctions locales dans un script - Matlab VS Octave

Dans les dernières versions de Matlab et Octave, on peut écrire les fonctions directement dans un script. En ajoutant des fonctions locales vous pouvez **éviter de créer et de gérer des fichiers de fonctions séparés**. Cependant, la syntaxe est différente entre Matlab et Octave.

Matlab, depuis la version R2016b. Les fonctions locales peuvent apparaître dans n'importe quel ordre mais doivent être placées **à la fin du fichier, après le code du script**. Voici un exemple :

```
x = 1:10;
n = length(x);
avg = mymean(x,n)
med = mymedian(x,n)

function a = mymean(v,n)
% MYMEAN Local function that calculates mean of array.
    a = sum(v)/n;
end

function m = mymedian(v,n)
% MYMEDIAN Local function that calculates median of array.
    w = sort(v);
    if rem(n,2) == 1
        m = w((n + 1)/2);
    else
        m = (w(n/2) + w(n/2 + 1))/2;
    end
end
```

Octave. Les fonctions locales peuvent apparaître dans n'importe quel ordre mais doivent être placées **avant le code du script et après l'instruction** `1;` Voici un exemple :

```
% Prevent Octave from thinking that this is a function file:
1;

function a = mymean(v,n)
  % MYMEAN Local function that calculates mean of array.
  a = sum(v)/n;
end

function m = mymedian(v,n)
% MYMEDIAN Local function that calculates median of array.
    w = sort(v);
    if rem(n,2) == 1
        m = w((n + 1)/2);
    else
        m = (w(n/2) + w(n/2 + 1))/2;
    end
end

x = 1:10;
n = length(x);
avg = mymean(x,n)
med = mymedian(x,n)
```

A.10.3. Fonctions anonymes (*lambda functions*)

Les fonctions anonymes permettent de définir une fonction directement dans le script, à condition que la fonction se compose d'une seule instruction. La syntaxe usuelle d'une fonction anonyme est

```
fun = @(arg1, arg2,...,argn) [expr] ; % crochets facultatifs si une seule expression
```

Nous utiliserons les fonctions anonymes surtout pour **écrire directement la fonction dans le fichier de script sans créer un fichier séparé en étant compatibles à la fois avec Matlab et avec Octave.**
Une application courante des fonctions anonymes consiste à définir une expression mathématique.

```
f = @(x) 2*x ; % equivaut a definir f(x)=2x
f(2) % on evalue f(2) et on obtient 4
```

Cela permet entre autre de calculer rapidement une solution approchée d'une équation :

```
% on veut resoudre x=cos(x)
f = @(x) x.^2-2 ; % on pose f(x)=x^2-2
% fsolve( fct dont on cherche un zero , un point pas trop eloigne de la solution )
fsolve( f, 1 )
fsolve( f,-1 )
```

La contrainte d'utiliser une seule instruction n'empêche pas de calculer plusieurs résultats (car on peut renvoyer un vecteur) ni d'écrire des boucles (grâce à l'utilisation des instructions pointées) ni des conditions (grâce à l'utilisation de masques, par exemple pour la définition d'une fonction par morceaux, comme on verra à la page 325).

A.11. Graphes de fonctions $\mathbb{R} \to \mathbb{R}$

Pour tracer le graphe d'une fonction $f : [a,b] \to \mathbb{R}$, il faut tout d'abord générer une liste de points x_i où évaluer la fonction f, puis la liste des valeurs $f(x_i)$ et enfin, avec la fonction `plot`, Octave reliera 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

- soit l'instruction `linspace(a,b,n)` qui construit la liste de n éléments

$$[a, a+h, a+2h, \ldots, b = a+nh] \quad \text{avec } h = \frac{b-a}{n-1}$$

```
x=linspace(1,5,5)
x =
   1 2 3 4 5
```

- soit l'instruction `[a:h:b]` qui construit la liste de $n = E(\frac{b-a}{h}) + 1$ éléments

$$[a, a+h, a+2h, \ldots, a+nh]$$

Dans ce cas, attention au dernier terme : b peut ne pas être pris en compte.

Voici un exemple avec une sinusoïde (en utilisant la fonction prédéfinie sin) :

```
x = linspace(-5,5,101); # x = [-5:0.1:5] 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 marges et sauvegarder dans différents formats.

Si la fonction n'est pas prédéfinie, il est bonne pratique de la définir pour qu'elle opère composante par composante lorsqu'on lui passe un vecteur ou une matrice. Les opérations `/`, `*` et `\^` agissant sur elle doivent être remplacées par les opérations point correspondantes `./`, `.*` et `\.^` qui opèrent composante par composante.
Par exemple, on se propose de tracer la fonction

$$f : [-2;2] \to \mathbb{R}$$
$$x \mapsto \frac{1}{1+x^2}$$

Suivant la façon de définir la fonction, on pourra utiliser l'une des trois méthodes suivantes.

Méthode 1. En utilisant une **fonction anonyme**.
Dans un script ou dans la `prompt` on écrit les instructions suivantes :

```
f=@(x)[1./(1+x.^2)] % declaration de la fonction
x=[-2:0.5:2];
y=f(x); % evaluation en plusieurs points
plot(x,y) % affichage des points (x_i,y_i)
```

Méthode 2. En utilisant une **fonction locale**.
Dans un script on écrit les instructions suivantes (attention : syntaxe différente selon qu'on utilise Matlab ou Octave) :

Matlab

```
x=[-2:0.5:2];
y=f(x); % evaluation en plusieurs points
plot(x,y) % affichage des points (x_i,y_i)

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

Octave

```
1;

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

x=[-2:0.5:2];
y=f(x); % evaluation en plusieurs points
plot(x,y) % affichage des points (x_i,y_i)
```

Méthode 3. En utilisant un **fichier fonction**.
On utilise deux fichiers :

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

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

3.2. Dans un script ou dans la `prompt` on écrit

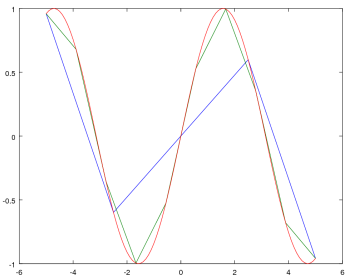
```
x=[-2:0.5:2];
y=f(x); % evaluation en plusieurs points
plot(x,y) % affichage des points (x_i,y_i)
```

A.11.1. Plusieurs courbes sur le même repère

On peut **tracer plusieurs courbes sur le même repère**. Par défaut, MATLAB/Octave efface la figure avant chaque commande de traçage `plot`. Vous pouvez tracer plusieurs lignes soit en écrivant plusieurs courbes dans une même instruction `plot` ou à l'aide de la commande de mise en attente `hold on`. Tant que vous n'utilisez pas de suspension (`hold off`) ou que vous ne fermez pas la fenêtre, tous les tracés apparaissent dans la fenêtre de la figure actuelle.

Par exemple, dans la figure suivante, on a tracé la même fonction : la courbe bleu correspond à la grille la plus grossière, la courbe rouge correspond à la grille la plus fine :

```
a = linspace(-5,5,5); % a = [-5,-3,-1,1,3,5]
fa = sin(a);
b = linspace(-5,5,10); % b = [-5,-4,-3,...,5]
fb = sin(b);
c = linspace(-5,5,101); % c = [-5,-4.9,-4.8,...,5]
fc = sin(c);
plot(a,fa,b,fb,c,fc)
% la derniere ligne peut etre remplacee par
% hold on
% plot(a,fa)
% plot(b,fb)
% plot(c,fc)
% hold off
```



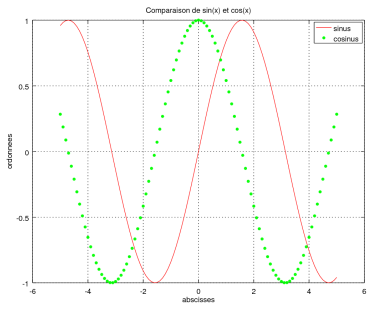
On peut spécifier la couleur et le type de trait, changer les étiquettes des axes, donner un titre, ajouter une grille, une légende etc.

Par exemple, dans le code ci-dessous `"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.

	linestyle=		color=		marker=
-	solid line	r	red	.	points
-	dashed line	g	green	,	pixel
:	dotted line	b	blue	o	filled circles
-.	dash-dot line	c	cyan	v	triangle down
		m	magenta	^	triangle up
		y	yellow	>	triangle right
		w	white	<	triangle left symbols
		k	black	*	star
				+	plus
				s	square
				p	pentagon
				x	x
				X	x filled
				d	thin diamond
				D	diamond

TABLE A.1. – Quelques options de `plot`

```
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
```

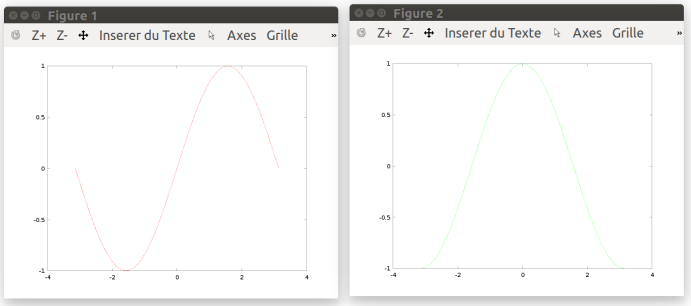


Voir la table A.1 et la documentation de `Matlab` pour connaître les autres options.

A.11.2. Plusieurs “fenêtres” graphiques

Avec `figure()` on génère une nouvelle fenêtres graphique :

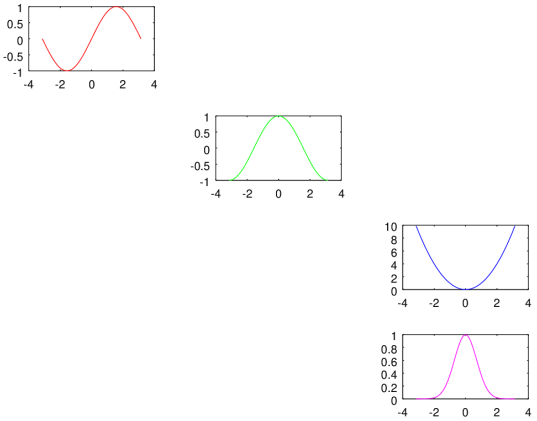
```
x = [-pi:0.05*pi:pi];
figure(1)
plot(x, sin(x), 'r')
figure(2)
plot(x, cos(x), 'g')
```



A.11.3. Plusieurs repères dans la même fenêtre

La fonction `subplot(x,y,z)` subdivise la fenêtre 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(4,3,1)
plot(x, sin(x), 'r')

subplot(4,3,5)
plot(x, cos(x), 'g')

subplot(4,3,9)
plot(x, x.*x, 'b')

subplot(4,3,12)
plot(x, exp(-x.*x), 'm')
```



EXEMPLE

Dans le code suivant on voit comment tracer plusieurs courbes dans un même graphique (avec légende), plusieurs graphe 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));
```

A.12. Structure conditionnelle

Supposons vouloir calculer la valeur $y = f(x)$ d'un nombre x selon la règle suivante :

$$f(x) = \begin{cases} x & \text{si } x \leq -5, \\ 100 & \text{si } -5 < x \leq 0, \\ x^2 & \text{si } 0 < x < 10, \\ x-2 & \text{sinon.} \end{cases}$$

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 complète 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.

Pour l'exemple donné, la fonction (vectorisée) peut s'écrire comme suit :

```
function y=f1(x)
   n=length(x);
   for i=1:n
      if x(i)<=-5
         y(i)=x(i);
      elseif x(i)<=0
         y(i)=100;
      elseif x(i)<10
         y(i)=x(i)^2;
      else
         y(i)=x(i)-2;
      endif
   endfor
endfunction
```

La même fonction peut aussi s'écrire comme suit :

```
f2 = @(x) [ x.*(x<=-5) + 100.*(x>-5).*(x<=0) + (x.^2).*(x>0).*(x<10) + (x-2).*(x>=10) ];
```

Pour vérifier qu'on a bien la même fonction on compare le graphe des deux fonctions :

```
xx=[-7:0.1:12];
yy1=f1(xx);
yy2=f2(xx);
plot(xx,yy1,'r',xx,yy2,'b.')
```

Voici un exemple pour établir si un nombre est positif

```
a=-1.5;

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

Voici un exemple pour établir si un nombre est compris entre deux valeurs :

```
x = 10;
minVal = 2;
maxVal = 6;

if (x >= minVal) && (x <= maxVal) % equivaut a (x >= minVal) .* (x <= maxVal)
    disp('Value within specified range.')
elseif (x > maxVal)
    disp('Value exceeds maximum value.')
else
    disp('Value is below minimum value.')
end
```

Voir aussi https://fr.mathworks.com/help/matlab/ref/if.html

Exemple

Étant donné deux matrices d'entrée $\mathbb{A}$ et $\mathbb{B}$, vérifier si on peut calculer le produit $\mathbb{AB}$. Si c'est le cas, créez une matrice $\mathbb{C}$ qui contient le produit $\mathbb{AB}$, sinon, $\mathbb{C}$ doit contenir une chaîne de caractère contenant un message d'erreur.

```
function C = in_prod(A,B)
  [rA,cA]=size(A);
  [rB,cB]=size(B);
  if cA==rB
      C = A*B;
  else
      C = "Have you checked the inner dimensions?"
  end
end

# TESTS
C=in_prod([1 2],[2;3])
C=in_prod(-5,100)
C=in_prod([1 2;3 4],[5;6])
C=in_prod([1 2 3; 4 5 6],[2 5;3 6])
```

A.13. Structures itératives

Les structures de répétition se classent en deux catégories : les *répétitions inconditionnelles* pour lesquelles le bloc d'instructions est à répéter un nombre donné de fois et les *répétitions conditionnelles* pour lesquelles le bloc d'instructions est à répéter autant de fois qu'une condition est vérifiée.

A.13.1. Répétition `for`

Lorsque l'on souhaite répéter un bloc d'instructions un nombre déterminé de fois, on peut utiliser un *compteur actif*, c'est-à-dire une variable qui compte le nombre de répétitions et conditionne la sortie de la boucle.
La syntaxe de la commande `for` est schématiquement

```
for var = expression
   instruction_1
   instruction_2
end
```

`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 $u_1 = u_2 = 1$:

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

```
n = 12;
u = [1,1];
for i = 3:n
   u = [ u , u(end)+u(end-1) ]; # concatenation
end
disp(u)
```

Dans les deux cas le résultat affiché est

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

Dans l'exemple suivant on calcul la somme des n premiers entiers (et on vérifie qu'on a bien $n(n+1)/2$) :

```
n=100;
s=0;
for i=1:n
   s += i;
end
s
n*(n+1)/2
```

Bien sur, dans ce cas il est préférable d'écrire

```
sum(1:100)
```

Il est possible d'imbriquer des boucles, c'est-à-dire que dans le bloc d'une boucle, on utilise une nouvelle boucle.

```
for x = [10,20,30,40,50] % for x = 10:10:50
   for y=[3,7]
      disp(x+y)
   end
end
```

Dans ce petit programme x vaut d'abord 10, y prend la valeur 3 puis la valeur 7 (le programme affiche donc d'abord 13, puis 17). Ensuite $x = 20$ et y vaut de nouveau 3 puis 7 (le programme affiche donc ensuite 23, puis 27).
Voir aussi https://fr.mathworks.com/help/matlab/ref/for.html

A.13.2. Boucle `while` : répétition conditionnelle

`While` est la traduction de "tant que...". Concrètement, la boucle s'exécutera tant qu'une condition est remplie (donc tant qu'elle renverra la valeur `1`). Le constructeur `while` a la forme générale suivante :

```
while condition
   instruction_1
   instruction_2
end
```

où `condition` représente des ensembles d'instructions dont la valeur est `1` ou `0`. Tant que la condition `condition` a la valeur `1`, on exécute les instructions `instruction_i`.

ATTENTION
Si la condition ne devient jamais fausse, le bloc d'instructions est répété indéfiniment et le programme ne se termine pas.

Voici un exemple pour créer la liste $\left[1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}\right]$:

```
nMax = 4;
n = 1;
a = [];
while n<=nMax
   a=[a,1/n]; # Append element to list
   n += 1;
end
disp(a)
```

Dans l'exemple suivant on calcul la somme des n premiers entiers tant que la somme ne dépasse pas 100 :

```
s=0;
n=0;
while s<100
   n += 1;
   s += n;
end
disp(n-1)
disp(s-n)
% En effet avec le dernier n on depasse 100
```

A.13.3. 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 récursive, 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.
Dans les exemples suivant, la fonction `tic` s'utilise avec la fonction `toc` pour mesurer le temps écoulé. La fonction `tic` enregistre l'heure actuelle et la fonction `toc` utilise la valeur enregistrée pour calculer le temps écoulé.

EXEMPLE
Quasiment toutes les fonctions prédéfinies sont vectorisées. Voici un exemple avec la fonction sin.

Le code

```
n = 100000;
xx = linspace(0,2*pi,n);
yy = zeros(length(xx));
tic
for i = 1:n
   yy(i) = sin(xx(i));
end;
toc
```

est significativement plus lent que

```
n = 100000;
xx = linspace(0,2*pi,n);
tic
yy = sin(xx);
toc
```

EXEMPLE
Pour calculer $\sum_{n=1}^{10000} \frac{1}{n^2}$, on peut utiliser les trois codes suivants, le deuxième étant significativement plus rapide :

```
n=1:10^7;

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

```
n=1:10^7;

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

```
n=1:10^7;

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

Un exemple de sorties obtenues :
- ⋆ avec le premier script : `Elapsed time is 12.7138 seconds.`
- ⋆ avec le deuxième script : `Elapsed time is 0.116321 seconds.`
- ⋆ avec le troisième script : `Elapsed time is 0.098047 seconds.`

A.14. 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

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

Les $n+1$ valeurs réels $a_0, a_1, \dots, a_n$ sont appelés les **coordonnées de p_n dans la base canonique**[3] de $\mathbb{R}_n[x]$ et on peut les stocker dans un vecteur **p** :

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

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 **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 $p(1) = a_n$, …, $p(n+1) = a_0$. Par exemple, pour construire le polynôme $p(x) = 2 - x + x^2$ nous écrirons

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

1. 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 **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])
   ```

2. Utilisée avec la commande `fplot`, la commande `polyval` permet de tracer le graphe de la fonction polynomiale sur un intervalle `[xmin, xmax]` 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])
   ```

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

3. 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)
```

4. 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 **r** est un vecteur contenant ses racines. Par exemple, pour définir le polynôme $p(x) = (x-1)(x+1)$ nous écrirons

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

5. 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
```

6. 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+3x^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
```

7. 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

```
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
```

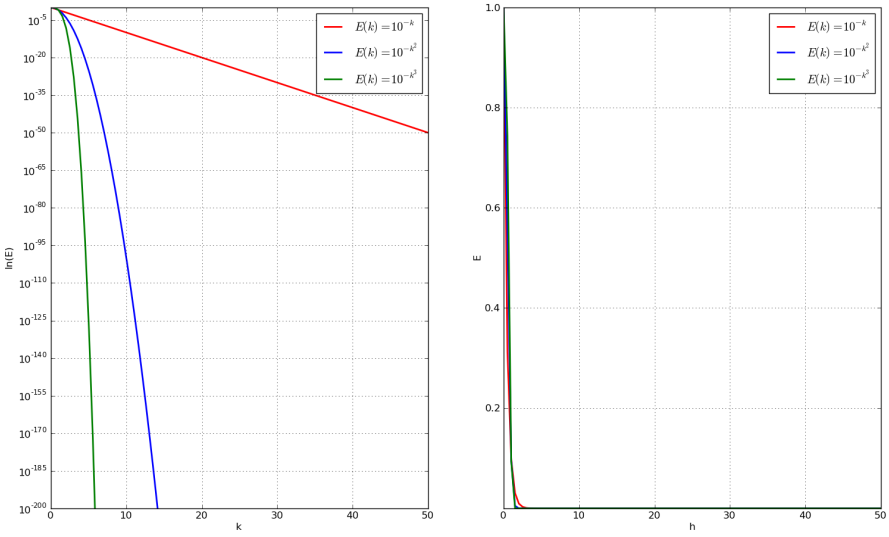


FIGURE A.1. – Échelle semi-logarithmique vs linéaire

9. La commande `polyint` permet de calculer le polynôme $\int_0^x p(t)\,\mathrm{d}t$ qui s'annule en 0 et qui est une primitive d'un polynôme p. La syntaxe est `polyint(p)`. Par exemple, pour calculer $\int_0^x p(t)\,\mathrm{d}t$ 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 $\int_0^3 p(t)\,\mathrm{d}t$ 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')
```

A.15. Exercices

Pensez à placer la commande `clear all` au début de vos scripts, de manière à nettoyer l'environnement de travail (cela effacera toutes les variables en mémoire). Vous pouvez aussi utiliser la commande `clc` pour nettoyer la fenêtre de commandes.

☆ Exercice A.1

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

① Somme et produit de matrices, calcul du déterminent 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 A.2 (Opérations élément par élément, produit scalaire, produit vectoriel)

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

```
clear all;
clc;
x = [1; 2; 3]
v = x.^2
b = sum(x)

p = x.^x

y = [4; 5; 6]
u = x.*y
w = sum(x.*y)
d = dot(x,y)
xTy = x'*y
yTx = y'*x

c = cross(x,y)
```

② 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`.

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 impaires inférieurs ou égaux à 500. Calculer la somme des nombres paires inférieurs ou égaux à 500.

Correction

① **x**, **y** et **v** sont des vecteurs-colonne 3×1

v est le vecteur tel que $v_i = x_i^2$ pour $i=1,2,3$: $\mathbf{v}=(1,4,9)^T$

$b = x_1+x_2+x_3 = 1+2+3 = 6$

p est le vecteur tel que $p_i = (x_i)^{x_i}$ pour $i=1,2,3$: $\mathbf{p}=(1,4,27)^T$

u est le vecteur tel que $u_i = x_i y_i$ pour $i=1,2,3$: $\mathbf{u}=(4,10,18)^T$

w, d, $\mathbf{x}^T\mathbf{y}$ et $\mathbf{y}^T\mathbf{x}$ sont quatre méthodes pour calculer le produit scalaire de **x** et **y** qui est égale à $\sum_i y_i x_i = y_1x_1+y_2x_2+y_3x_3 = 4\times 1+5\times 2+6\times 3 = 32$

c est le vecteur obtenu par le produit vectoriel de **x** et **y** : $\mathbf{c}=(-3,6,-3)^T$

②
```
x=[pi/6 pi/4 pi/3]
s=sin(x)
c=cos(x)
t=s./c
tan(x) % on verifie qu'on a le bon resultat
```

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

```
n=500
sum([1:n])
n*(n+1)/2
```

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

```
n=500
sum([1:n].^2)
n*(n+1)*(2*n+1)/6
```

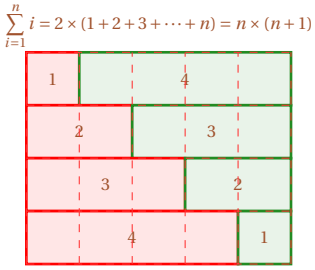
$\sum_{i=1}^{N}(2i-1) = N^2$

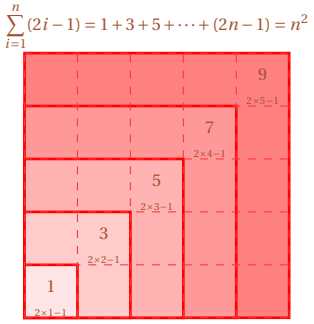
```
n=500
imp=sum(1:2:n)
N=length(1:2:n);
N^2
```

$\sum_{i=1}^{N}(2i) = N(N+1)$

```
n=500
pair=sum(2:2:n)
N=length(2:2:n);
N*(N+1)
```

Preuves sans mot





Remarque

Quelque somme remarquable :

1. $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$.

2. $\sum_{i=1}^{n}(2i-1)=n^2$. En effet, on remarque qu'il s'agit d'une somme télescopique :

$$\sum_{i=1}^{n}(2i-1)=\sum_{i=1}^{n}(i^2-(i-1)^2)=(1^2-0^2)+(2^2-1^2)+(3^2-2^2)+\ldots(n^2-(n-1)^2)=n^2.$$

3. $\sum_{i=1}^{n}(2i)=2\sum_{i=1}^{n}i=n(n+1)$.
4. $\sum_{i=1}^{n}i^2=\dfrac{n(n+1)(2n+1)}{6}$. En effet,

$$\sum_{i=1}^{n}(i+1)^3=\sum_{j=1}^{n}j^3-1+(n+1)^3, \qquad \sum_{i=1}^{n}(i+1)^3=\sum_{i=1}^{n}i^3+3\sum_{i=1}^{n}i^2+3\sum_{i=1}^{n}i+\sum_{i=1}^{n}1$$

donc

$$\begin{aligned}3\sum_{i=1}^{n}i^2&=\sum_{j=1}^{n}j^3-1+(n+1)^3-\sum_{i=1}^{n}i^3-3\sum_{i=1}^{n}i-\sum_{i=1}^{n}1\\&=-1+(n+1)^3-3\frac{n(n+1)}{2}-n=\frac{(n+1)}{2}\left(2(n+1)^2-3n-2\right)=\frac{n(n+1)(2n+1)}{2}.\end{aligned}$$

5. $\sum_{i=1}^{n}i^3=\left(\dfrac{n(n+1)}{2}\right)^2$. En effet,

$$\sum_{i=1}^{n}(i+1)^4=\sum_{j=1}^{n}j^4-1+(n+1)^4, \qquad \sum_{i=1}^{n}(i+1)^4=\sum_{i=1}^{n}i^4+4\sum_{i=1}^{n}i^3+6\sum_{i=1}^{n}i^2+4\sum_{i=1}^{n}i+\sum_{i=1}^{n}1$$

donc

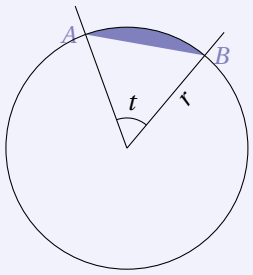
$$\begin{aligned}4\sum_{i=1}^{n}i^3&=\sum_{j=1}^{n}j^4-1+(n+1)^4-\sum_{i=1}^{n}i^4-4\sum_{i=1}^{n}i^3-6\sum_{i=1}^{n}i^2-4\sum_{i=1}^{n}i-\sum_{i=1}^{n}1\\&=-1+(n+1)^4-6\frac{n(n+1)(2n+1)}{6}-4\frac{n(n+1)}{2}-n=(n+1)\left((n+1)^3-n(2n+1)-2n-1\right)\\&=(n+1)^2\left((n+1)^2-(2n+1)\right)=n^2(n+1)^2.\end{aligned}$$

☆ Exercice A.3

On se propose ici d'utiliser Octave pour résoudre graphiquement des équations.

Considérons un cercle de rayon r. Si nous traçons un angle t (mesuré en radians) à partir du centre du cercle, les deux rayons formant cet angle coupent le cercle en A et B. Nous appelons a l'aire délimitée par la corde et l'arc AB (en bleu sur le dessin). Cette aire est donnée par $a=\dfrac{r^2}{2}(t-\sin(t))$. Pour un cercle donné (c'est à dire un rayon donné), nous choisissons une aire (partie en bleu) a. Quelle valeur de l'angle t permet d'obtenir l'aire choisie? Autrement dit, connaissant a et r, nous voulons déterminer t solution de l'équation

$$\frac{2a}{r^2}=t-\sin(t).$$



1. Résoudre graphiquement l'équation en traçant les courbes correspondant aux membres gauche et droit de l'équation (pour $a=4$ et $r=2$). Quelle valeur de t est solution de l'équation?
2. Comment faire pour obtenir une valeur plus précise du résultat?

Correction

```
t=[0:pi/180:2*pi];
rhs=t-sin(t);
a=4;
r=2;
lhs=2*a/(r^2)*ones(size(t));
plot(t,rhs,t,lhs), grid
```

Le graphe nous dit que la solution est entre 2 et 3. On peut calculer une solution approchée comme suit :

```
fsolve( @(x) 2*a/(r^2)-(x-sin(x)) , 2.5)
```

☆ Exercice A.4 (Vectorisation)

Réécrire les codes suivants sans utiliser de boucles :

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

②
```
n=length(b);
for i = 1:n-1
   a(i) = b(i+1) - b(i);
end
```

③
```
n=length(a);
for i = 1:n-1
   if (a(i) > 5)
      a(i) -= 20
   end
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
```

☆ Exercice A.5 (Vectorisation)

Considérons la fonction $f : \mathbb{R} \to \mathbb{R}$ définie par

$$f(x) = \begin{cases} 0 & \text{si } x < 0, \\ x & \text{si } 0 \le x < 1, \\ 2-x & \text{si } 1 \le x \le 2, \\ 0 & \text{si } x > 2. \end{cases}$$

Écrire et afficher cette fonction en deux manières différentes :

1. avec une instruction conditionnelle du type `if ... elseif ... else` non vectorisée,
2. avec une instruction conditionnelle vectorisée.

Correction

La fonction non vectorisée peut s'écrire comme suit :

```
function y=f1(x)
   n=length(x);
   for i=1:n
      if x(i)<0
           y(i)=0;
      elseif x(i)<1
         y(i)=x(i);
      elseif x(i)<=2
         y(i)=2-x(i);
      else
         y(i)=0;
      endif
   endfor
endfunction
```

La fonction vectorisée peut s'écrire comme suit :

```
f2 = @(x) [ 0*(x<0) + x.*(x>=0).*(x<1) + (2-x).*(x>=1).*(x<2) + 0.*(x>2) ];
```

On compare le graphe des deux fonctions :

```
xx=[-1:0.1:3];
yy1=f1(xx);
yy2=f2(xx);
plot(xx,yy1,'r',xx,yy2,'b.')
```

☆ Exercice A.6 (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

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

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

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

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. $\sum_{i=1}^{n} w_i x_i y_i = \mathbf{x}\mathbb{D}_{\mathbf{w}}\mathbf{y}^T$ 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 A.7
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 $1,\dots,10$, *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 `A*B` calcule le produit $\mathbb{A}\mathbb{B} = (\sum_{k=1}^{3} a_{ik}b_{kj})_{1\le i,j\le 3}$, `A.*B` calcule la matrice $\mathbb{C} = (a_{ij}b_{ij})_{1\le i,j\le 3}$.

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

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

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

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

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

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

```
>> 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 $\mathbb{A}$ qui contient juste les produits, et la matrice $\mathbb{B}$ qui contient aussi l'entête des lignes et colonnes :

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

☆ Exercice A.8 (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 $\mathbb{A}$ 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 $\mathbb{V}^{-1}\mathbb{A}\mathbb{V} = \mathbb{D}$ 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.

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 ensuite 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 $\mathbb{A}$ 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 $\mathbb{Z}$ 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 $1,2,\ldots,n^2$ 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 = [v1;v2;v3;v4]` et utiliser le fait que les colonnes sont linéairement indépendants ssi le déterminant de $\mathbb{A}$ 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 Soit n et m fixés.

```
A(1:n,:)=ones(n,1)*[1:m]
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 A.9 (Construction de matrices, vectorisation, `script` et `function`)**

1. Dans un **fichier** `zorro.m` écrire une `function` appelée `zorro` qui prend en entrée un entier $n \in \mathbb{N}^*$ et renvoi la matrice carrée $\mathbb{Z}$ 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

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

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

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
```

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

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

☆ Exercice A.10 (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 A.11 (Boucle `for`)

Il est possible de calculer les premières décimales de π avec l'aide du hasard. On considère un carré de coté 1 et un cercle de rayon 1 centré à l'origine :

Si on divise l'aire de la portion de disque par celle du carré on trouve $\frac{\pi}{4}$. Si on tiré au hasard dans le carré, on a une probabilité de $\frac{\pi}{4}$ que le point soit dans la portion de disque. On considère l'algorithme suivant pour approcher π : on génère N couples $\{(x_k, y_k)\}_{k=1}^{N}$ de nombres aléatoires dans l'intervalle $[0,1]$, puis on calcule le nombre $m \leq N$ de ceux qui se trouvent dans le premier quart du cercle unité. π est la limite de la suite $4m/N$ lorsque $N \to +\infty$. 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=10^4
```

```
# On tire au hasard N points [x,y] dans [0,1[ x [0,1[
[xx,yy]=rand(N,2);
# Nombre de tirs dans le disque
m=sum( xx.^2+yy.^2<=1 );
myPi=4*m/N
err=abs(pi-myPi)
```

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)<= 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 peut réécrire le tout comme une fonction anonyme :

```
format long
myPi = @(N) [ 4*sum( rand(N,1).^2+rand(N,1).^2<=1 )/N ];
err=abs(pi-myPi(10^4)) % test
```

On exécute maintenant le programme pour différentes valeurs de N. Plus N est grand, meilleure est l'approximation de π. 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).

```
format long
myPi = @(N) [ 4*sum( rand(N,1).^2+rand(N,1).^2<=1 )/N ];
NN=100:100:10000;
for i=1:length(NN)
    err(i)=abs(pi-myPi(NN(i)));
end
plot(NN,err)
```

Cette méthode n'est pas très efficace, il faut beaucoup de tirs pour obtenir le deux premières décimales de π.

☆ Exercice A.12 (`function`)

Comme π vérifie

$$\pi = \lim_{N\to+\infty} \sum_{n=0}^{N} 16^{-n}\left(\frac{4}{8n+1} - \frac{2}{8n+4} - \frac{1}{8n+5} - \frac{1}{8n+6}\right)$$

on peut calculer une approximation de π en sommant les N premiers termes, pour N assez grand.

- ⋆ Écrire une fonction pour calculer les sommes partielles de cette série (*i.e.* pour $n = 0 \dots N$ avec N donné en paramètre).
- ⋆ Pour quelles valeurs de N obtient-on une approximation de π aussi précise que celle fournie par la variable `pi` ?

Correction
Pour répondre à la question on peut utiliser le script suivant :

```
format long

piapproche = @(v) sum( ( 4./(8*v+1) - 2./(8*v+4) - 1./(8*v+5) - 1./(8*v+6) ).*(1/16).^v );

N=0;
while abs(pi-piapproche([0:N]))>0
    N+=1;
end
N
piapproche([0:N])
pi
```

Pour $n = 10$ on obtient une approximation de π 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 π.

☆ Exercice A.13

Un dispositif fournit un signal $s(t) = A\sin(2\pi t + \varphi)$ avec A et φ inconnus. On mesure le signal à deux instants (en ms) : $s(0.5) = -1.76789123$ et $s(0.6) = -2.469394443$. On posera $\alpha = A\cos(\varphi)$ et $\beta = A\sin(\varphi)$.

1. Écrire et résoudre le système d'inconnues α et β. En déduire A et φ.

2. Tracer le signal et montrer qu'il passe par les points mesurés.

Correction
Rappel : $\sin(a+b) = \sin(a)\cos(b) + \cos(a)\sin(b)$ ainsi

$$s(t) = A\sin(2\pi t + \varphi) = A\big(\sin(2\pi t)\cos(\varphi) + \cos(2\pi t)\sin(\varphi)\big) = \alpha\sin(2\pi t) + \beta\cos(2\pi t).$$

On doit donc résoudre le système linéaire :

$$\begin{cases} \alpha\sin(\pi) + \beta\cos(\pi) = -1.76789123 \\ \alpha\sin\left(\frac{6}{5}\pi\right) + \beta\cos\left(\frac{6}{5}\pi\right) = -2.469394443 \end{cases}$$

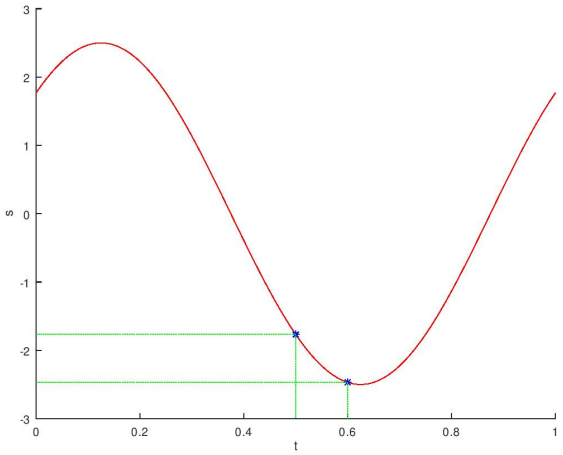
qu'on écriture matricielle s'écrit

$$\begin{pmatrix} \sin(\pi) & \cos(\pi) \\ \sin\left(\frac{6}{5}\pi\right) & \cos\left(\frac{6}{5}\pi\right) \end{pmatrix}\begin{pmatrix}\alpha\\\beta\end{pmatrix} = \begin{pmatrix} -1.76789123 \\ -2.469394443 \end{pmatrix} \quad i.e. \quad \begin{pmatrix} 0 & -1 \\ \sin\left(\frac{6}{5}\pi\right) & \cos\left(\frac{6}{5}\pi\right) \end{pmatrix}\begin{pmatrix}\alpha\\\beta\end{pmatrix} = \begin{pmatrix} -1.76789123 \\ -2.469394443 \end{pmatrix}$$

```
xx=[sin(2*pi*0.5) , cos(2*pi*0.5) ; sin(2*pi*0.6) , cos(2*pi*0.6)]\[-1.76789123;-2.469394443]
alpha=xx(1)
beta=xx(2)
```

Puisqu'on trouve $\alpha = \beta$, alors $\cos(\varphi) = \sin(\varphi)$, *i.e.* $\varphi = \frac{\pi}{4}$ et $A = \sqrt{2}\alpha$ avec $\alpha \approx 1.7679$:

```
s=@(t)[sqrt(2)*alpha*sin(2*pi*t+pi/4)];
tt=[0:0.01:1];
hold on
plot(tt,s(tt),'r-') % la courbe
plot([0.5 0.6],[-1.76789123;-2.469394443],'b*') % les deux points
plot([0 0.5 0.5], [-1.76789123 -1.76789123 -3], 'g:') % trait pointille
plot([0 0.6 0.6], [-2.469394443 -2.469394443 -3], 'g:') % trait pointille
xlabel('t')
ylabel('s')
hold off
print("signal.jpg") % sauvagarde de la figure en .jpg
```



☆ Exercice A.14 (Résolution graphique d'une équation)

Soit la fonction

$$f : [-10, 10] \to \mathbb{R}$$
$$x \mapsto \frac{x^3\cos(x) + x^2 - x + 1}{x^4 - \sqrt{3}x^2 + 127}$$

1. Tracer le graphe de la fonction f en utilisant seulement les valeurs de $f(x)$ lorsque la variable x prend successivement les valeurs $-10, -9.2, -8.4, \ldots, 8.4, 9.2, 10$ (*i.e.* avec un pas 0.8).

2. Apparemment, l'équation $f(x) = 0$ a une solution α voisine de 2. En utilisant le zoom, proposer une valeur approchée de α.
3. Tracer de nouveau le graphe de f en faisant varier x avec un pas de 0.05. Ce nouveau graphe amène-t-il à corriger la valeur de α proposée?
4. Demander à Octave d'approcher α (fonction `fsolve`).

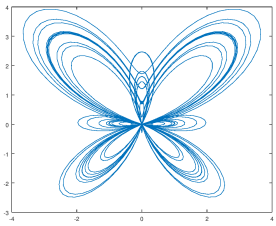
Correction
En utilisant un pas de 0.8 il semblerai que $\alpha = 1.89$. En utilisant un pas de 0.05 il semblerai que $\alpha = 1.965$. En utilisant la fonction `fsolve` on trouve $\alpha = 1.9629$.

```
clear all; clc;
f = @(x) [(x.^3 .*cos(x)+x.^2-x+1) ./ (x.^4-sqrt(3)*x.^2+127)] ;

subplot(1,2,1)
xx=[-10:0.8:10];
plot(xx,f(xx),'r-')
grid()

subplot(1,2,2)
xx=[-10:0.05:10];
plot(xx,f(xx),'r-')
grid()

fsolve(f,1.9)
```

☆ Exercice A.15 (Courbe parametrée)
Tracer la courbe papillon ($t \in [0;100]$) :

$$\begin{cases} x(t) = \sin(t)\left(e^{\cos(t)} - 2\cos(4t) - \sin^5\left(\frac{t}{12}\right)\right) \\ y(t) = \cos(t)\left(e^{\cos(t)} - 2\cos(4t) - \sin^5\left(\frac{t}{12}\right)\right) \end{cases}$$

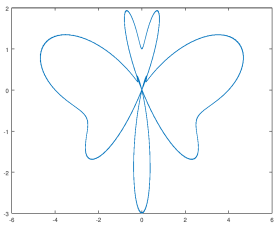
```
x=@(t) [ sin(t).*( exp(cos(t))-2*cos(4*t)-(sin(t/12)).^5 ) ];
y=@(t) [ cos(t).*( exp(cos(t))-2*cos(4*t)-(sin(t/12)).^5 ) ];
tt=[0:0.05:100];
plot(x(tt),y(tt))
```



☆ Exercice A.16 (Équation polaire)
Tracer la courbe papillon ($t \in [0;100]$) :

$$\begin{cases} x(t) = r(t)\cos(t) \\ y(t) = r(t)\sin(t) \end{cases} \quad \text{avec } r(t) = \sin(7t) - 1 - 3\cos(2t).$$

```
r=@(t) [sin(7*t)-1-3*cos(2*t)];
x=@(t) [ r(t).*cos(t) ];
y=@(t) [ r(t).*sin(t) ];
tt=[0:0.05:100];
plot(x(tt),y(tt))
```



☆ Exercice A.17 (`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 accepte que r et h soient des listes de nombres et renvoyer un tableau. Si r est un vecteur de n nombres et h de m nombres alors v sera une matrice $\mathbb{A}$ de dimension $n \times m$ telle que $a_{ij} = v(r_i, h_j)$.

Correction
On écrit la fonction soit avec `function` soit avec une fonction anonyme, puis on valide la fonction avec un test dont on connaît le résultat :

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

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.

A.15.1. Suites

☆ Exercice A.18
Soit $(u_n)_{n\in\mathbb{N}}$ la suite définie par $u_n = (0.7)^{3n}$. Quel est le plus petit n tel que $u_n < 10^{-4}$? (Calculer d'abord analytiquement pui vérifier numériquement le résultat.)

Correction
$u_n = \left(\left(\frac{7}{10}\right)^3\right)^n = \left(\frac{343}{1000}\right)^n$. Il s'agit d'une suite géométrique de raison $0 < q < 1$: elle est donc décroissante. On a $u_n < 10^{-4}$ ssi $\left(\frac{343}{1000}\right)^n < 10^{-4}$ ssi $\log_{10}\left(\frac{343}{1000}\right)^n < -4$ ssi $n > -\frac{4}{\log_{10}\left(\frac{343}{1000}\right)} = -\frac{4}{\log_{10}(343)-\log_{10}(10^3)} = \frac{4}{3-\log_{10}(343)} \simeq \frac{4}{0.5} = 8$. La valeur cherchée est donc $n = 9$.
Vérifions nos calculs :

```
n=0
u=1
while u>=1.e-4
   n+=1
   u=(0.7)^(3*n)
end
```

☆ Exercice A.19
On achète un ordinateur portable à 430 e. On estime qu'une fois sorti du magasin sa valeur u_n en euro après n mois est donnée par la formule

$$u_n = 40 + 300 \times (0.95)^n.$$

1. Que vaut l'ordinateur à la sortie du magasin ?
2. Que vaut-il après un an de l'achat ?
3. À long terme, à quel prix peut-on espérer revendre cet ordinateur ?
4. Déterminer le mois à partir duquel l'ordinateur aura une valeur inférieure à 100 e.

Correction
1. À la sortie du magasin $u_0 = 340$
2. Après un an de l'achat on a $u_{12} = 40 + 300 \times (0.95)^{12} = 202.11$
3. À long terme, on peut espérer revendre cet ordinateur à $\lim_{n\to+\infty} u_n = 40$.
4. À partir du 32-ème mois l'ordinateur aura une valeur inférieure à 100 e car :

$$40 + 300 \times (0.95)^n < 100 \iff (0.95)^n < \frac{100-40}{300} = \frac{1}{5} = 5^{-1} \iff n\ln(0.95) < -\ln(5) \iff n > -\frac{\ln(5)}{\ln(0.95)} \simeq 31.377$$

Vérifions nos calculs :

```
nn=[0];
uu=[430];
for n=1:50
   nn=[nn,n];
   uu=[uu, 40+300*(0.95)^n];
end
plot(nn,uu,'*-',nn,100*ones(size(nn)),'-')
indice=max(find(uu>100))
printf(strcat("u_",num2str(nn(indice)),"=",num2str(uu(indice)),"\n"))
printf(strcat("u_",num2str(nn(indice+1)),"=",num2str(uu(indice+1)),"\n"))
plot(nn,uu,'*-',nn,100*ones(size(nn)),'-',[nn(indice),nn(indice)],[0,uu(indice)],[nn(indice+1),nn(indice
    +1)],[0,uu(indice+1)])
```

☆ Exercice A.20 (Suite de FIBONACCI)

La suite de FIBONACCI est une suite d'entiers dans laquelle chaque terme est la somme des deux termes qui le précèdent. Elle commence généralement par les termes 0 et 1 (parfois 1 et 1). Elle doit son nom à Leonardo FIBONACCI, un mathématicien italien du XIII^{e} siècle qui, dans un problème récréatif posé dans un de ses ouvrages, le *Liber Abaci*, décrit la croissance d'une population de lapins :

> «Un homme met un couple de lapins dans un lieu isolé de tous les côtés par un mur. Combien de couples obtient-on en un an si chaque couple engendre tous les mois un nouveau couple à compter du troisième mois de son existence?»

Le problème de FIBONACCI est à l'origine de la suite dont le n-ième terme correspond au nombre de paires de lapins au n-ème mois. Dans cette population (idéale), on suppose que :

- au (début du) premier mois, il y a juste une paire de lapereaux;
- les lapereaux ne procréent qu'à partir du (début du) troisième mois;
- chaque (début de) mois, toute paire susceptible de procréer engendre effectivement une nouvelle paire de lapereaux;
- les lapins ne meurent jamais (donc la suite de FIBONACCI est strictement croissante).

Notons F_n le nombre de couples de lapins au début du mois n. Jusqu'à la fin du deuxième mois, la population se limite à un couple (ce qu'on note $F_1 = F_2 = 1$). Dès le début du troisième mois, le couple de lapins a deux mois et il engendre un autre couple de lapins; on note alors $F_3 = 2$. Plaçons-nous maintenant au mois n et cherchons à exprimer ce qu'il en sera deux mois plus tard, soit au mois $n+2$: F_{n+2} désigne la somme des couples de lapins au mois $n+1$ et des couples nouvellement engendrés. Or, n'engendrent au mois $(n+2)$ que les couples pubères, c'est-à-dire ceux qui existent deux mois auparavant. On a donc, pour tout entier n strictement positif, $F_{n+2} = F_{n+1} + F_n$. On choisit alors de poser $F_0 = 0$, de manière que cette équation soit encore vérifiée pour $n = 0$. On obtient ainsi la forme récurrente de la suite de FIBONACCI : chaque terme de cette suite est la somme des deux termes précédents :

$$\begin{cases} F_0 = 0, \\ F_1 = 1, \\ F_{n+2} = F_{n+1} + F_n \end{cases}$$

On souhaite comparer les temps CPU pour calculer le 30-ème élément F_{30} de la suite de FIBONACCI suivant la méthode choisie :

Boucle calcul de F_n à l'aide d'une boucle;

Écriture matricielle calcul de F_n en exploitant la relation matricielle

$$\begin{bmatrix} F_n \\ F_{n-1} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} F_{n-1} \\ F_{n-2} \end{bmatrix}$$

Fonction récursive calcul de F_n à l'ide d'une fonction récursive

Expression fonctionnelle calcul de F_n en exploitant une expression fonctionnelle de la suite, c'est-à-dire une expression telle que le calcul de F_n pour une valeur de n donnée ne présuppose la connaissance d'aucune autre valeur de n.

Correction

Boucle On utilise l'écriture $F_{n+2} = F_{n+1} + F_n$ avec pour valeurs initiales $F_1 = F_2 = 1$. À chaque étape on peut stocker toutes les valeurs de la suite ou juste les trois dernières.

Écriture matricielle En exploitant la relation matricielle on obtient par récurrence

$$\begin{bmatrix} F_n \\ F_{n-1} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} F_{n-1} \\ F_{n-2} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}^2 \begin{bmatrix} F_{n-2} \\ F_{n-3} \end{bmatrix} = \cdots = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}^n \begin{bmatrix} F_1 \\ F_0 \end{bmatrix}$$

Expression fonctionnelle Comme la suite de FIBONACCI est linéaire d'ordre deux, on peut écrire son équation caractéristique. On obtient une équation du second degré $x^2 - x - 1 = 0$ qui a pour solutions $x_1 = \varphi = \frac{1+\sqrt{5}}{2}$ (le nombre d'or) et $x_2 = 1 - \varphi = \frac{1-\sqrt{5}}{2}$. Il en résulte que $F_n = \alpha\varphi^n + \beta(1-\varphi)^n$ où α et β sont deux constantes à déterminer à partir de F_0 et F_1. On a $\alpha + \beta = 0$ et $(\alpha - \beta)\varphi + \beta = 1$ ce qui donne $\alpha = -\beta = 1/\sqrt{5}$. On trouve alors l'expression générale de la suite de FIBONACCI (appelée formule de BINET) :

$$F_n = \frac{1}{\sqrt{5}}\left(\varphi^n - (1-\varphi)^n\right).$$

Si on calcule la limite du rapport de deux nombres consécutifs de la suite de FIBONACCI on trouve le nombre d'or :

$$\frac{F_{n+1}}{F_n} = \frac{\varphi^{n+1} - (1-\varphi)^{n+1}}{\varphi^n - (1-\varphi)^n} = \varphi \frac{1 - \left(\frac{1-\varphi}{\varphi}\right)^{n+1}}{1 - \left(\frac{1-\varphi}{\varphi}\right)^n} \xrightarrow[n\to\infty]{} \varphi$$

car $1 < \varphi < 2$ et donc $-1 < \frac{1-\varphi}{\varphi} < 1$.

```
clear all; clc;
n = 30;

# Boucle 1
printf('Boucle, on stocke toute la suite\n')
tic
F = ones(1,n);
for i = 3:n
   F(i) = F(i-1)+F(i-2);
end
F(end)
toc

# Boucle 2
printf('\nBoucle, on stocke juste 3 terms de la
    suite\n')
tic
a=1;
b=1;
for i = 3:n
   c = b+a;
   a=b;
   b=c;
end
c
toc

# Matricielle
printf('\nCalcul matriciel\n')
tic
fibmat(n)
toc

# Recoursive
printf('\nCalcul recursif\n')
tic
fibrec(n)
toc

# Fonctionnelle
sprintf('\nCalcul fonctionnel\n')
tic
#phi=fsolve(@(x)(x^2-x-1),1);
phi=(1+sqrt(5))/2;
Fn=(phi^n-(1-phi)^n)/sqrt(5)
toc
```

☆ Exercice A.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 366ème personne aura forcément une date d'anniversaire commune avec une personne déjà présente. [a]

Maintenant, passons à une question moins évidente : combien faut il réunir de personne 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, fatalement, 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 ;
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

a. 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.

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 A.22 (Conjecture de Syracuse)

Considérons la suite récurrente

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

En faisant des tests numérique on remarque que la suite obtenue tombe toujours sur 1 peut importe l'entier choisit [a] au départ. La conjecture de Syracuse affirme que, peu importe le nombre de départ choisi, la suite ainsi construite atteint le chiffre 1 (et donc boucle sur le cycle 4, 2, 1). Cet énoncé porte le nom de «Conjecture» et non de théorème, car ce résultat n'a pas (encore) été démontré pour tous les nombres entiers. En 2004, la conjecture a été "juste" vérifiée pour tous les nombres inférieurs à 2^{64}.

1. Écrire un script qui, pour une valeur de $u_1 \in]1; 10^6]$ 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 *durée 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 la plus grande durée de vol ou la plus grandes altitude maximale. On notera que, même en partant de nombre peu élevés, il est possible d'obtenir des altitudes très hautes. Vérifiez que, en partant de 27, elle atteint une altitude maximale de 9232 et une durée de vol de 111. Au contraire, on peut prendre des nombres très grands et voir leur altitude chuter de manière vertigineuse sans jamais voler plus haut que le point de départ. Faire le calcul en partant de 10^6.
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 $3n+1$). Vous pouvez lire l'article de vulgarisation https://automaths.blog/2017/06/20/la-conjecture-de-syracuse/amp/

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

Correction

```
function u = mysuite(u_init)
    u=[u_init];
    while u(end)~=1
      if rem(u(end),2)==0
         u=[u,u(end)/2];
      else
         u=[u,3*u(end)+1];
      end
   end
end
```

```
N=60;

L=[]; M=[]; F=[];

for n=[2:N]
    U=mysuite(n);
    L=[L,length(U)];
    M=[M,max(U(:))];
    F=[F,M(end)/n];
end

subplot(1,3,1)
maxL=max(L);
indicemaxL=find(L==maxL,1)+1
plot(Uinit,L,'b-',[0,indicemaxL,indicemaxL],[maxL,maxL,0],'r:')
title(strcat( ["Duree de vol:\n max=" num2str(maxL) " obtenu avec u_1=" num2str(indicemaxL)] ))
grid()

subplot(1,3,2)
maxM=max(M);
indicemaxM=find(M==maxM,1)+1
plot(Uinit,M,'b-',[0,indicemaxM,indicemaxM],[maxM,maxM,0],'r:')
title(strcat( ["Altitude maximale:\n max=" num2str(maxM) " obtenue avec u_1=" num2str(indicemaxM)] ))
grid()

subplot(1,3,3)
maxF=max(F);
indicemaxF=find(F==maxF,1)+1
plot(Uinit,F,'b-',[0,indicemaxF,indicemaxF],[maxF,maxF,0],'r:')
title(strcat( ["Facteur d'expansion:\n max=" num2str(maxF) " obtenu avec u_1=" num2str(indicemaxF)] ))
grid()
```

☆ Exercice A.23 (Fractales)

La répétition de transformations permet de tracer des figures géométriques appelées fractales.
Considérons la suite définie par récurrence suivante :

$$\begin{pmatrix} x \\ y \end{pmatrix}_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \begin{pmatrix} x \\ y \end{pmatrix}_{n+1} = \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}_n + \begin{pmatrix} q_1 \\ q_2 \end{pmatrix}_n$$

Pour chaque cas, tracer l'ensemble de points de coordonnées (x_n, y_n).

Exemple-1 on pose $\mathbb{M} = \frac{1}{2}\begin{pmatrix}1 & -1\\ 1 & 1\end{pmatrix}$ et $\mathbf{q} = \begin{pmatrix}1\\ -3\end{pmatrix}$,

Dragon de Heighway on choisit au hasard et de façon équiprobable $\alpha \in [0;1[$ puis on pose $\mathbb{M} = (\alpha < \frac{1}{2})\mathbb{M}_1 + (\alpha \geq \frac{1}{2})\mathbb{M}_2$ et $\mathbf{q} = (\alpha < \frac{1}{2})\mathbf{q}_1 + (\alpha \geq \frac{1}{2})\mathbf{q}_2$ avec $\mathbb{M}_1 = \frac{1}{2}\begin{pmatrix}1 & -1\\ 1 & 1\end{pmatrix}$, $\mathbb{M}_2 = \frac{1}{2}\begin{pmatrix}-1 & -1\\ 1 & -1\end{pmatrix}$ et $\mathbf{q}_1 = \begin{pmatrix}0\\ 0\end{pmatrix}$, $\mathbf{q}_2 = \begin{pmatrix}1\\ 0\end{pmatrix}$ avec $n = 0, \ldots, 10000$.

Fougère de Barnsley on choisit au hasard et de façon équiprobable $\alpha \in [0;1[$ puis on pose $\mathbb{M} = (\alpha < \frac{1}{100})\mathbb{M}_1 + (\frac{1}{100} \leq \alpha < \frac{86}{100})\mathbb{M}_2 + \frac{86}{100} \leq \alpha < \frac{93}{100})\mathbb{M}_3 + (\alpha \geq \frac{93}{100})\mathbb{M}_4$ et $\mathbf{q} = (\alpha < \frac{1}{100})\mathbf{q}_1 + (\frac{1}{100} \leq \alpha < \frac{86}{100})\mathbf{q}_2 + \frac{86}{100} \leq \alpha < \frac{93}{100})\mathbf{q}_3 + (\alpha \geq \frac{93}{100})\mathbf{q}_4$ avec $\mathbb{M}_1 = \frac{1}{100}\begin{pmatrix}0 & 0\\ 0 & 16\end{pmatrix}$, $\mathbb{M}_2 = \frac{1}{100}\begin{pmatrix}85 & 4\\ -4 & 85\end{pmatrix}$, $\mathbb{M}_3 = \frac{1}{100}\begin{pmatrix}20 & -26\\ 23 & 22\end{pmatrix}$, $\mathbb{M}_4 = \frac{1}{100}\begin{pmatrix}-15 & 28\\ 26 & 24\end{pmatrix}$ et $\mathbf{q}_1 = \begin{pmatrix}0\\ 0\end{pmatrix}$, $\mathbf{q}_2 = \begin{pmatrix}0\\ 1.6\end{pmatrix}$, $\mathbf{q}_3 = \mathbf{q}_2$, $\mathbf{q}_4 = \begin{pmatrix}0\\ 0.44\end{pmatrix}$ avec $n = 0, \ldots, 10000$.

Arbre on choisit au hasard et de façon équiprobable $\alpha \in [0;1[$ puis on pose $\mathbb{M} = (\alpha < \frac{1}{3})\mathbb{M}_1 + (\frac{1}{3} \leq \alpha < \frac{2}{3})\mathbb{M}_2 + (\alpha \geq \frac{2}{3})\mathbb{M}_3$ et $\mathbf{q} = (\alpha < \frac{1}{3})\mathbf{q}_1 + (\frac{1}{3} \leq \alpha < \frac{2}{3})\mathbf{q}_2 + (\alpha \geq \frac{2}{3})\mathbf{q}_3$ avec $\mathbb{M}_1 = \frac{1}{200}\begin{pmatrix}0 & 0\\ 0 & 51\end{pmatrix}$, $\mathbb{M}_2 = \frac{6}{8}\begin{pmatrix}\cos(\vartheta) & -\sin(\vartheta)\\ \sin(\vartheta) & \cos(\vartheta)\end{pmatrix}$, $\mathbb{M}_3 = \frac{1}{8}\begin{pmatrix}5\cos(\psi) & -6\sin(\psi)\\ 5\sin(\psi) & 6\cos(\psi)\end{pmatrix}$ et $\mathbf{q}_1 = \frac{1}{2}\begin{pmatrix}1\\ 0\end{pmatrix}$, $\mathbf{q}_2 = \begin{pmatrix}\frac{1}{2} - \frac{3}{8}\cos(\vartheta)\\ \frac{51}{200} - \frac{3}{8}\sin(\vartheta)\end{pmatrix}$, $\mathbf{q}_3 = \begin{pmatrix}\frac{1}{2} - \frac{5}{16}\cos(\psi)\\ \frac{153}{1000} - \frac{5}{16}\sin(\psi)\end{pmatrix}$ avec $n = 0, \ldots, 100$, $\vartheta = -\frac{\pi}{8}$ et $\psi = \frac{\pi}{5}$.

Source http://www.ac-grenoble.fr/maths/PM/Ressources/568/TP_fractales_Python.pdf

Correction

```
clc;
clear all;
cas=4;

% p et q vecteurs colonne, M matrice carree
transform = @(p,M,q) M*p+q;

% MAIN
n=1000*(cas==1)+10000*(cas==2)+10000*(cas==3)+100*(cas==4);
A = zeros(n,2);
for i=2:n
    if cas==1 % Exemple1
        M=[1,-1;1,1]/2;
        q=[1;-3];
    elseif cas==2 % Dragon
        M1=[1,-1;1,1]/2;
        q1=[0;0];
        M2=[-1,-1;1,-1]/2;
        q2=[1;0];
        alpha=rand();
        M=(alpha<0.5)*M1+(alpha>=0.5)*M2;
        q=(alpha<0.5)*q1+(alpha>=0.5)*q2;
    elseif cas==3 % Fougere
        M1=[0,0;0,0.16];
        q1=[0;0];
        M2=[0.85,0.04;-0.04,0.85];
        q2=[0;1.6];
        M3=[0.2,-0.26;0.23,0.22];
        q3=[0;1.6];
        M4=[-0.15,0.28;0.26,0.24];
        q4=[0;0.44];
        alpha=rand();
        M=(alpha<0.01)*M1+(alpha>=0.01)*(alpha<0.86)*M2+(alpha>=0.86)*(alpha<0.93)*M3+(alpha>=0.93)*M4;
        q=(alpha<0.01)*q1+(alpha>=0.01)*(alpha<0.86)*q2+(alpha>=0.86)*(alpha<0.93)*q3+(alpha>=0.93)*q4;
    else % arbre
        M1=[0,0;0,51/200];
        q1=[0.5;0];
```

```
        theta=-pi/8;
        M2=6/8*[cos(theta),-sin(theta);sin(theta),cos(theta)];
        q2=[1/2-3/8*cos(theta);51/200-3/8*sin(theta)];
        psi=pi/5;
        M3=1/8*[5*cos(psi),-6*sin(psi);5*sin(psi),6*cos(psi)];
        q3=[1/2-5/16*cos(psi);153/1000-5/16*sin(psi)];
        alpha=rand();
        M=(alpha<1/3)*M1+(alpha>=1/3)*(alpha<2/3)*M2+(alpha>=2/3)*(alpha<0.93)*M3;
        q=(alpha<1/3)*q1+(alpha>=1/3)*(alpha<2/3)*q2+(alpha>=2/3)*(alpha<0.93)*q3;
    end
    A(i,:)=transform(A(i-1,:)',M,q);
end
plot(A(:,1),A(:,2),'o','MarkerSize',8)
```

A.15.2. Systèmes linéaires

☆ Exercice A.24 (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 utilise la fonction *linsolve*.

Méthode 4. 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]

disp('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.

disp('Methode 2')
x = A\b
norm(A*x-b)
```

```
disp('Methode 3')
x=linsolve(A,b)
norm(A*x-b)

disp('Methode 4')
Aaug=[A b]
RRAaug=rref(Aaug)
x=RRAaug(:,4)
norm(A*x-b)

disp('Methode 5')
[L,U,P]=lu(A);
y=L\(P*b);
x=U\y
norm(A*x-b)
```

☆ Exercice A.25 (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)
```

☆ Exercice A.26

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

$$\mathbb{A} = \begin{pmatrix} 1 & 0 & 3 \\ 2 & 2 & 2 \\ 3 & 6 & 4 \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} 4 \\ 6 \\ 13 \end{pmatrix}$$

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 qua Ax=b
printf(strcat("||Ax-b||=",num2str(norm(A*x-b)),"\n"))
```

A.15.3. Traitement mathématique des images numérique

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

- **Acquisition d'une image (= numérisation)** : dans la figure ci dessous on a à gauche une image réelle et à droite sa version numérisée. La numérisation d'une image réelle comporte une perte d'information due d'une part à l'échantillonnage (procédé de discrétisation spatiale d'une image consistant à projeter sur une grille régulière, *i.e.* en un nombre fini de points, une image analogique continue en lui associant une valeur unique), d'autre part à la quantification (nombre finis de nuances = limitation du nombre de valeurs différentes que peut prendre un point).
- **Les pixels d'une image** : une image numérique en niveaux de gris (*grayscale* image en anglais) est un tableau de valeurs $\mathbb{A}$. Chaque case de ce tableau, qui stocke une valeur a_{ij}, se nomme un pixel (*PICTure ELement*).

  En notant n le nombre de lignes et p le nombre de colonnes du tableau, on manipule ainsi un tableau de $n \times p$ 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.

  On peut définir une fonction comme suit :

$$\begin{aligned} g\colon [\![1:n]\!] \times [\![1:p]\!] &\to [\![0:255]\!] \\ (i,j) &\mapsto g(i,j) = a_{ij} \end{aligned}$$

- **La taille d'une image** : la taille d'une image est le nombre de pixel. Les dimensions d'une image sont la largeur (=nombre de colonnes du tableau) et la hauteur (=nombre de lignes du tableau).
- **La résolution d'une image** : la résolution d'une image est le nombre de pixel par unité de longueur. En générale, on utilise des "pixel par pouce" ou "point par pouce" (ppp), on anglais on dit *dot per inch* (`dpi`) : n `dpi` = n pixel pour un puce = n pixel pour 2.54 cm.

Les fonctions Octave utiles pour gérer les images sont les suivantes :

- `image` : affiche une image (objet graphique Image) ;
- `imagesc` ou `imshow` : affiche une image (objet graphique Image) avec interpolation des couleurs ;
- `imread` : lit une image d'un fichier (formats standards) ;
- `imwrite` : écrit une image dans fichier (formats standards) ;
- `imfinfo` : extrait des informations d'un fichier (formats standards) ;
- `print` : exporte une image (formats standards).

L'exemple suivant montre une visualisation d'un tableau carré avec $n = p = 512$, ce qui représente $512 \times 512 = 2^{18} = 262\,144$ pixels. [4] Dans ce cas, nous ne construisons pas la matrice à la main, mais nous allons lire un fichier image et l'importer comme une matrice dans Octave.

4. Les appareils photos numériques peuvent enregistrer des images beaucoup plus grandes, avec plusieurs millions de pixels.

Pour **transformer une image en une matrice** il suffit d'indiquer dans un script : [5]

```
A=imread('lena512.bmp');
```

Avec la fonction `imread` Octave transforme un fichier image (ici de type `.bmp`) en une matrice dont chaque coefficient est un flottant si l'image est en noir et blanc.
On peut **voir l'image** avec :

```
imshow(uint8(A));
```

ou avec

```
imagesc(A)
axis image
```

On peut **sauvegarder l'image** avec :

```
imwrite(uint8(A),'monimage.jpg','jpg');
```

FIGURE A.2. – Léna (original)
http://www.lenna.org

Pour connaître la **dimension de la matrice** (*i.e.* de l'image) il suffira d'écrire

```
[row,col]=size(A)
```

On a une matrice de taille 512×512.
Pour connaître l'**intervalle des valeurs de la matrice** on écrira

```
pp=min(A(:))
pg=max(A(:))
```

Les niveaux de gris sont compris entre 25 et 245.
Attention : lorsque l'on charge une image, la matrice correspondante est de type `uint8` (codage sur 8 bits non signés), cela signifie que ses éléments (qui représentent les niveaux de gris) sont dans l'intervalle d'entiers $[\![0-255]\!]$). Octave peut lire des images codées sur 8, 16, 24 ou 32 bits. Mais le stockage et l'affichage de ces données ne peut être fait qu'avec trois types de variables :

- ⋆ le type `uint8` (entier non signé de 8 bits) de plage $[0;255]$;
- ⋆ le type `uint16` (entier non signé de 16 bits) de plage $[0;65535]$;
- ⋆ le type `double` (réel 64 bits) de plage $[0;1]$.

Dans le code ci-dessous, on a $a = 155$ mais a est de type `uint8` donc, lorsqu'on calcule a^2, Octave plafonne à 255 :

```
A=imread('lena512.bmp');
a=A(10,10);
disp(a) % output 155
disp(155^2) % output 24025
disp(a^2) % output 255
a=uint16(a);
disp(a^2) % output 24025
```

☆ Exercice A.27 (Manipulations élémentaires)

1. Que se passe-t-il lorsqu'on exécute le script suivant?

```
clear all

A=imread('lena512.bmp');
colormap(gray(256));

subplot(1,2,1)
 imshow(uint8(A));
 title ( "Original" );
subplot(1,2,2)
 B=A';
 imshow(uint8(B));
 title ( "Transposee" );
 imwrite(uint8(B),'exotransposee.jpg','jpg');
```

2. Que se passe-t-il lorsqu'on exécute le script suivant?

5. Cette image, "Lena", est une image digitale fétiche des chercheurs en traitement d'images. Elle a été scannée en 1973 dans un exemplaire de Playboy et elle est toujours utilisée pour vérifier la validité des algorithmes de traitement ou de compression d'images. On la trouve sur le site de l'University of Southern California http://sipi.usc.edu/database/.

```
clear all

A=imread('lena.jpg');
[row,col]=size(A);
```

```
for i=1:col
   A=[A(:,2:col) A(:,1)];
   imshow(A,gray(256));
   pause(0.001);
end
```

3. En utilisant une manipulation élémentaire de la matrice (sans faire de boucles et sans utiliser de fonctions) obtenir les images de la figure A.3.

(a) Original (b) Flip vertical (c) Flip horizontale (d) Flip double (e) Hstack

(f) Vstack (g) Zoom (h) Effacer (i) Bord carré (j) Bord cercle

FIGURE A.3. – Manipulations élémentaires

Correction

1. 
Originale


Transposée

2. À chaque étape on enlève la première colonne et on la concatène comme dernière colonne. Le résultat donne un "film" dans lequel l'image "sort" à gauche et "rentre" à droite.
3. On pourra se baser sur le canevas suivant :

```
clear all
A=imread('lena512.bmp');
B= ; % a completer
subplot(1,2,1)
 imshow(uint8(A));
 title ( "Originale" );
subplot(1,2,2)
 imshow(uint8(B));
 title ( "Transformee" );
```

3.1. Flip vertical
```
B=A(end:-1:1,:);
```

3.2. Flip horizontale
```
B=A(:,end:-1:1);
```

3.3. Flip double
```
B=A(end:-1:1,end:-1:1);
```

3.4. Hstack
```
B=[A(:,end/2:end)A(:,end:-1:end/2)];
```

3.5. Vstack `B=[ A(end/2:end,:); A(end:-1:end/2,:)];`

3.6. Zoom `B=A(300:400,250:350);`

3.7. Effacer `B=A; B(250:300,220:375)=0;`

3.8. Bord carré `B=A; [r,c]=size(A); B([1:20,r-20:r],:)=0; B(:,[1:20,c-20:c])=0;`

3.9. Bord cercle

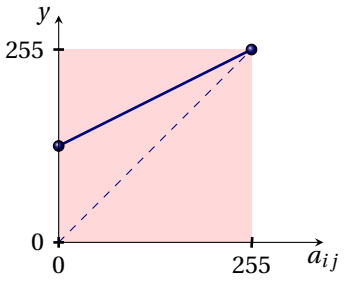
```
[r,c]=size(A);
B=255*ones(r,c);
for i=1:r
  for j=1:c
    if (i-r/2)^2+(j-c/2)^2 <= (r/2)^2
      B(i,j)=A(i,j);
    end
  end
end
```

On peut améliorer visuellement l'image en modifiant la valeur de chaque pixel par une fonction qu'on appliquera à tous les pixels. Pour cela, on pourra se baser sur le canevas suivant :

```
clear all
A=double(imread('lena512.bmp')); % utiliser double pour avoir des calculs precis

f=@(g) .... ; % fonction vectorisee a completer

B=f(A);
subplot(1,3,1)
 plot([0:255], f([0:255]), 'b-',[0:255], [0:255], 'r--'); % f et identite
 axis([0 255 0 255],"square");
 title('f vs identite')
subplot(1,3,2)
 imshow(uint8(A));
 title ( "Originale" );
subplot(1,3,3)
 imshow(uint8(B));
 title ( "Transformee" );
```

Par exemple, la fonction suivante, appliquée à chaque pixel d'une image, éclaircira l'image :

$$f\colon [0;255] \to [0;255]$$
$$a_{ij} \mapsto \frac{a_{ij}+255}{2}$$



Dans le canevas, il suffit de définir

```
f = @(g) (g+255)/2;
```

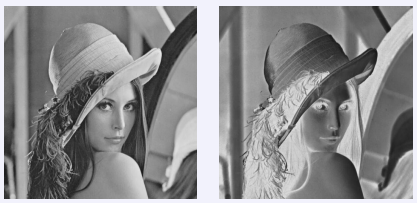
et on obtient

⭐ Exercice A.28 (Luminosité et contraste)

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

$$f : [0;255] \to [0;255]$$
$$g \mapsto 255 - g$$

Appliquer cette transformation pour obtenir l'image A.4b.

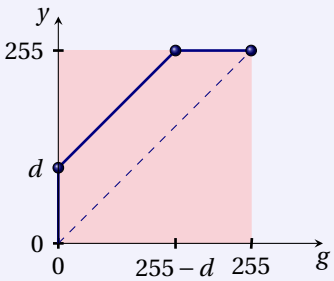


(a) Originale (b) Négatif

FIGURE A.4. – Négatif

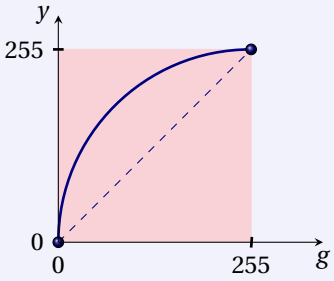
2. Appliquer la transformation ci-dessous pour augmenter la luminosité en ajoutant la valeur fixe $d = 50$ à tous les niveaux de gris et obtenir l'image A.5b.

$$f : [0;255] \to [0;255]$$
$$g \mapsto \begin{cases} g+d & \text{si } g \leq 255-d, \\ 255 & \text{sinon.} \end{cases}$$



3. Il est très mauvais d'augmenter ainsi la luminosité : avec un décalage de d, il n'existera plus aucun point entre 0 et d et les points ayant une valeur supérieure à $255 - d$ deviendront des points parfaitement blancs, puisque la valeur maximale possible est 255. La nouvelle image contient des zones brûlées.

Plutôt que d'utiliser la fonction donnée, il vaut mieux utiliser une fonction bijective de forte croissance au voisinage de 0 et de très faible croissance au voisinage de 255, comme sur le graphe ci-contre.
Appliquer cette transformation pour obtenir l'image A.5c.

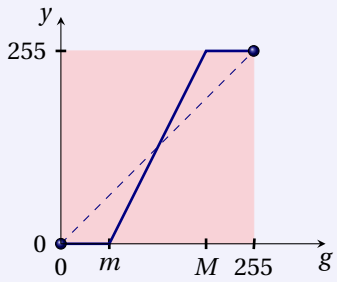


(a) Originale (b) Première méthode (c) Deuxième méthode

FIGURE A.5. – Modification de la luminosité

4. On peut composer avec une fonction pour *étaler* l'histogramme sur [0;255] (*Histogram Stretching*). Par exemple, si $m = \min(\mathbb{A})$ et $M = \max(\mathbb{A})$, on peut utiliser la fonction :

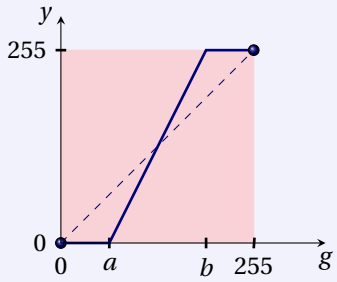
$$f : [0;255] \to [0;255]$$
$$g \mapsto \frac{255-0}{M-m}(g-m)+0$$



Appliquer cette transformation pour obtenir l'image A.6b.

5. On peut augmenter le contraste (*Contrast Stretching*) en "mappant" par une fonction linéaire par morceaux.

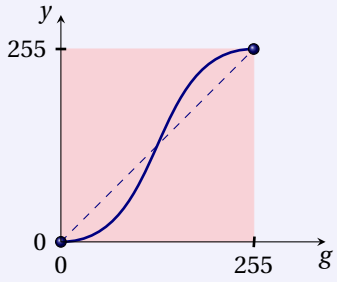
$$f : [0;255] \to [0;255]$$
$$g \mapsto \begin{cases} 0 & \text{si } g \le a \\ \frac{255-0}{b-a}(g-a)+0 & \text{si } a < g < b \\ 255 & \text{si } g \ge b \end{cases}$$



Appliquer cette transformation avec $a = 64$ et $b = 192$ pour obtenir l'image A.6c.

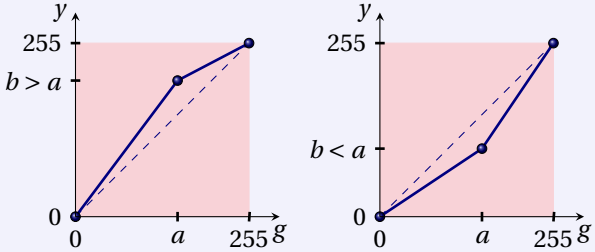
Un cas particulier s'obtient lorsque $a = b$ (*Contrast Thresholding* ou *seuillage*). Appliquer cette transformation pour obtenir l'image A.6d

Avec cette transformation, les points les plus blancs auront une valeur égale à b et il n'existera plus aucun point entre b et 255. De même, les points ayant une valeur comprise entre 0 et a deviendront noirs, puisque la valeur minimale est 0. Il y aura donc là perte d'informations. Ainsi il est plus judicieux d'adoucir la courbe :



6. On peut rehausser le contraste en "mappant" par une fonction linéaire par morceaux

$$f : [0;255] \to [0;255]$$
$$g \mapsto \begin{cases} \frac{b}{a}g & \text{si } g \le a \\ \frac{(255-b)g+255(b-a)}{255-a} & \text{si } g \ge a \end{cases}$$

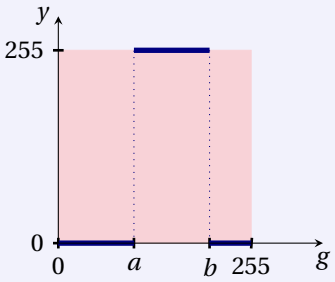


Appliquer cette transformation avec $a = 100$ et $b = 200$ (dilatation de la dynamique des zones claires) et comparer avec $a = 100$ et $b = 50$ (dilatation de la dynamique des zones sombres). Appliquer ces transformations pour

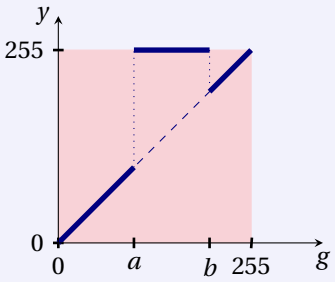
obtenir les images A.6e et A.6f.

7. On pourrait vouloir mettre en avant certains niveaux de gris dans un certain intervalle. Cette approche peut prendre deux formes : le *gray level slicing without background* et le *gray level slicing with background*. Dans la première approche, une grande valeur est donnée aux pixels dont le niveau de gris appartient à la plage choisie et les autres sont remplacés par 0. Dans la deuxième approche on ne modifie que les pixels à mettre en valeur. Cela correspond aux deux fonctions suivantes :

$$f\colon [0;255] \to [0;255]$$
$$g \mapsto \begin{cases} 0 & \text{si } g \le a \text{ ou } g \ge b \\ 255 & \text{si } a \le g \ge b \end{cases}$$



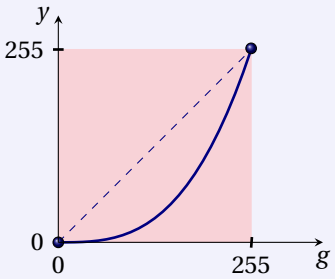
$$f\colon [0;255] \to [0;255]$$
$$g \mapsto \begin{cases} g & \text{si } g \le a \text{ ou } g \ge b \\ 255 & \text{si } a \le g \ge b \end{cases}$$



Appliquer ces deux transformations avec $a = 100$ et $b = 155$ pour obtenir les images A.6g et A.6h .

8. On peut modifier le contraste en “mappant” par une fonction croissante plus rapidement ou plus lentement que la fonction identité $i\colon [0;255] \to [0;255]$, $i(g) = g$. Par exemple, la fonction suivante modifie la caractéristique de gamma (plus clair si $0 < a < 1$, plus foncé si $a > 1$).

$$f\colon [0;255] \to [0;255]$$
$$g \mapsto 255\left(\frac{g}{255}\right)^a$$



Appliquer cette transformation avec $a = 3$ et $a = 1/3$ pour obtenir les images A.6i et A.6j.

(a) Originale (b) Histogram Stretching (c) Contrast Stretching (d) Contrast Thresholding (e) Dilatation de la dynamique des zones claires

(f) Dilatation de la dynamique des zones sombres (g) Gray level slicing without background (h) Gray level slicing with background (i) Non linear (j) Non linear

FIGURE A.6. – Modification du contraste

Correction

1. Négatif
   ```
   f = @(g)255-g;
   ```

2. *Histogram Stretching :*
   ```
   M=max(A(:)); m=min(A(:)); f = @(g)255*(g-m)/(M-m);
   ```

3. *Contrast Stretching* :
   ```
   a=64; b=192; f = @(g)(g<=a)*0 + (g>=b)*255 + (g>a).*(g<b).*( 255/(b-a)*(g-a));
   ```
   Si $a = m$ et $b = M$ on retrouve l'*Histogram Stretching*.

   Contrast Thresholding :
   ```
   a=128; f = @(g)(g<=a)*0 + (g>=a)*255;
   ```

   Exemple de fonction en S (à modifier pour qu'elle passe en $(0,0)$ et en $(255,255)$) :
   ```
   f = @(g)255/pi*atan((g-255/2)/10)+255/2;
   ```

4. Dilatation de la dynamique des zones claires :
   ```
   a=100; b=200; f = @(g)(g<=a)*(b/a).*g + (g>a).*( ((255-b)*g+255*(b-a))/(255-a));
   ```

   Dilatation de la dynamique des zones sombres :
   ```
   a=100; b=50; f = @(g)(g<=a)*(b/a).*g + (g>a).*( ((255-b)*g+255*(b-a))/(255-a));
   ```

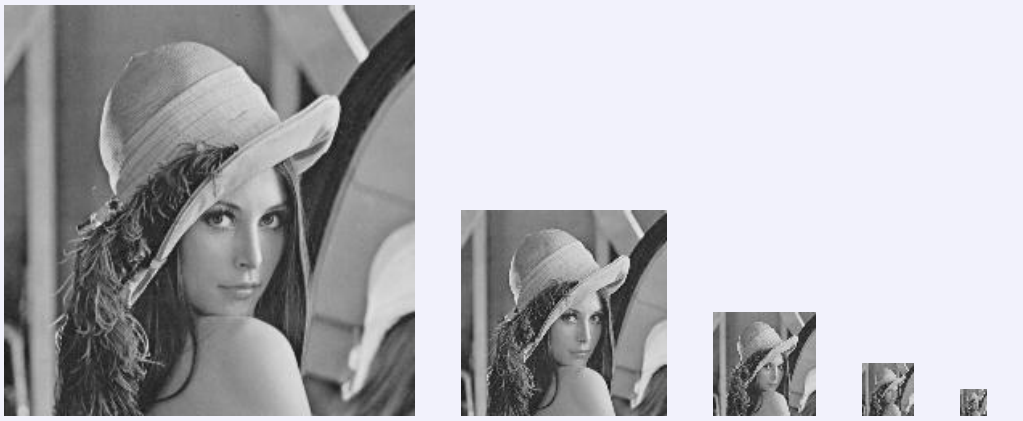
5. *Gray level slicing without background :*
   ```
   a=85; b=170; f = @(g)(g<=a)*0 + (g>=b)*0 + (g>a).*(g<b)*255 ;
   ```

   Gray level slicing with background :
   ```
   a=85; b=170; f = @(g)(g<=a).*g + (g>=b).*g + (g>a).*(g<b)*255 ;
   ```

6. *Non linear :*
   ```
   a=3; f = @(g)255*(g/255).^a;
   a=1/3; f = @(g)255*(g/255).^a;
   ```

☆ Exercice A.29 (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^k et une colonne sur 2^k ce qui donne une matrice $2^{9-k} \times 2^{9-k}$. Appliquer cette transformation pour obtenir l'une des images suivantes :



(a) Originale ($k = 1$) (b) $k = 2$ (c) $k = 3$ (d) $k = 4$ (e) $k = 5$

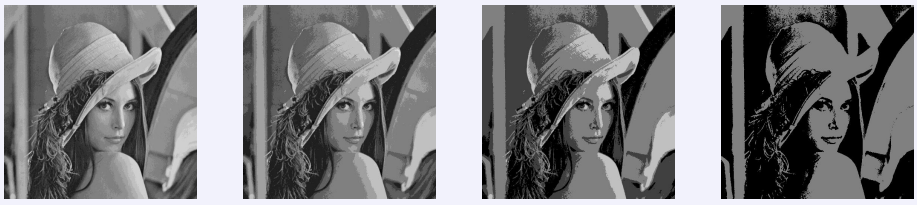
FIGURE A.7. – Résolution

Correction

```
clear all

A=double(imread('lena512.bmp'));
colormap(gray(256));

[row,col]=size(A)

for k=1:8
  subplot(1,8,k)
   E=A(1:2^k:row,1:2^k:col);
   imshow(uint8(E));
   title(['k=' num2str(k)]);
   file=strcat('exo2E', num2str(k) ,'.jpg')
   imwrite(uint8(E),file,'jpg');
end
```

☆ Exercice A.30 (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 :



(a) 16 niveaux de gris ($k = 4$) (b) 8 niveaux de gris ($k = 5$) (c) 4 niveaux de gris ($k = 6$) (d) 2 niveaux de gris ($k = 7$)

FIGURE A.8. – Quantification

Correction

```
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(2,2,1)
    imshow(uint8(A));
    title ( "Original" );
    subplot(2,2,3)
    hist(A(:),0:255);
```

☆ Exercice A.31 (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

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

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

Appliquer 100 fois cette transformation à la matrice $\mathbb{A}$ pour obtenir l'image A.9c. [a]
Appliquer ensuite la détection des bords (normalisée) à l'image A.9c pour obtenir l'image A.9d.

(a) Norme du gradient (b) Norme normalisée (c) Floutage (d) Gradient normalisée après floutage

FIGURE A.9. – Détection des bords et Floutage

a. 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

Correction

```
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 ramene a [0;255]
   m=min(min(G));
   M=max(max(G));
   G=255/(M-m).*(G-m);
   AA=G;
end

subplot(2,2,1)
 imshow(uint8(A));
 title ( "Original" );
subplot(2,2,3)
 hist(A(:),0:255);
subplot(2,2,2)
 imshow(uint8(G));
 title ( "Floutage" );
 imwrite(uint8(G),'exo5.jpg','jpg');
subplot(2,2,4)
 hist(G(:),0:255);
```

☆ Exercice A.32 (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 :

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

avec

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

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

Appliquer cette transformation suivie de la transformation en négatif pour obtenir l'image A.9a. 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 A.9b. Cela correspond à la transformation

$$f : [m;M] \to [0;255]$$
$$g \mapsto \frac{255}{M-m}(g-m)$$

Correction

```
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(2,3,1)
 imshow(uint8(A));
 title ( "Original" );
subplot(2,3,4)
 hist(A(:),0:255);
subplot(2,3,2)
 imshow(uint8(G));
 title ( "Gradient (negatif)" );
 imwrite(uint8(G),'exo41.jpg','jpg');
subplot(2,3,5)
 hist(G(:),0:255);
subplot(2,3,3)
 imshow(uint8(Gn));
 title ( "Gradient normalise (negatif)" );
 imwrite(uint8(Gn),'exo42.jpg','jpg');
subplot(2,3,6)
 hist(Gn(:),0:255);
```

☆ Exercice A.33 (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 512×512. 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 A.10. – 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 $i = s+1, \ldots, r$. 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 $\mathbb{A}^T\mathbb{A}$.

$$\mathbb{A}\mathbb{A}^T = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 2 \end{pmatrix} \qquad \mathbb{A}^T\mathbb{A} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

Valeurs propres :

$$\lambda_1 = 2 > \lambda_2 = \lambda_3 = 0 \qquad \lambda_1 = 2 > \lambda_2 = 0$$

Vecteurs propres unitaires :

$$\mathbb{U} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \qquad \mathbb{V} = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}$$

Donc

$$\mathbb{A} = \mathbb{U}\mathbb{S}\mathbb{V}^T = \underbrace{\begin{pmatrix} \mathbf{u}_1 & \ldots & \mathbf{u}_r & \mathbf{u}_{r+1} & \ldots & \mathbf{u}_n \end{pmatrix}}_{\in \mathbb{R}^{n \times n}} \underbrace{\begin{pmatrix} \sigma_1 & & & & \\ & \ddots & & & \\ & & \sigma_r & & \\ & & & 0 & \\ & & & & \ddots \\ & & & & & 0 \end{pmatrix}}_{\in \mathbb{R}^{n \times p}} \underbrace{\begin{pmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_r^T \\ \mathbf{v}_{r+1}^T \\ \vdots \\ \mathbf{v}_p^T \end{pmatrix}}_{\in \mathbb{R}^{p \times p}}$$

$$= \underbrace{\begin{pmatrix} \mathbf{u}_1 & \ldots & \mathbf{u}_r \end{pmatrix}}_{\in \mathbb{R}^{n \times r}} \underbrace{\begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{pmatrix}}_{\in \mathbb{R}^{r \times r}} \underbrace{\begin{pmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_r^T \end{pmatrix}}_{\in \mathbb{R}^{r \times p}} = \sum_{i=1}^{r} \sigma_i \underbrace{\mathbf{u}_i \times \mathbf{v}_i^T}_{\in \mathbb{R}^{r \times r}}$$

devient

$$\mathbb{A} = \left(\begin{array}{cc|c} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{array}\right) \left(\begin{array}{cc} \sqrt{2} & 0 \\ 0 & 0 \\ \hline 0 & 0 \end{array}\right) \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}$$

$$\overset{r=2}{=}\begin{pmatrix}0&1\\0&0\\1&0\end{pmatrix}\begin{pmatrix}\sqrt{2}&0\\0&0\end{pmatrix}\begin{pmatrix}1/\sqrt{2}&1/\sqrt{2}\\1/\sqrt{2}&-1/\sqrt{2}\end{pmatrix}$$
$$=\sqrt{2}\begin{pmatrix}0&0\\0&0\\1/\sqrt{2}&1/\sqrt{2}\end{pmatrix}+0\begin{pmatrix}1/\sqrt{2}&-1/\sqrt{2}\\0&0\\0&0\end{pmatrix}\overset{s=1}{=}\sqrt{2}\begin{pmatrix}0&0\\0&0\\1/\sqrt{2}&1/\sqrt{2}\end{pmatrix}$$

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 A.34 (Traitement mathématique des images numériques - compression par SVD)

1. Tester la compression avec $s=10$ et $s=100$ pour obtenir les images A.11 ainsi que la carte des erreurs.
2. Calculer $\max\left\{s\in[0;r]\;\middle|\;s<\frac{np}{n+p+1}\right\}$ qui est la limite en dessous de laquelle le stockage de la décomposition SVD permet des économies de mémoire par rapport au stockage de l'image initiale.
3. Supposons que la précision d'Octave soit de l'ordre de 3 chiffres significatifs. Alors les valeurs singulières significatives doivent avoir un rapport de moins de 10^{-3} avec la valeur maximale σ_1, les autres étant considérées comme «erronées». Calculer le nombre de valeurs singulières «significatives», *i.e.* la plus grande valeur de s telle que $\frac{\sigma_i}{\sigma_1}<10^{-3}$ pour $i=s+1,\dots,r$.

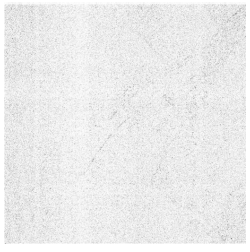
(a) Original $s=n=p=512$

(b) $\min\left\{s\in[0;r]\;\middle|\;\frac{\sigma_s}{\sigma_1}<10^{-5}\right\}=275$

(c) $\max\left\{s\in[0;r]\;\middle|\;s<\frac{np}{n+p+1}\right\}=255$

(d) $s=100$

(e) $s=10$



(f) $\min\left\{s\in[0;r]\;\middle|\;\frac{\sigma_s}{\sigma_1}<10^{-5}\right\}=275$

(g) $\max\left\{s\in[0;r]\;\middle|\;s<\frac{np}{n+p+1}\right\}=255$

(h) $s=100$

(i) $s=10$

FIGURE A.11. – SVD - exercice A.34

Correction

```
clear all

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

colormap(gray(256));

subplot(5,2,1)
imshow(uint8(A));
s=min(row,col);
title ( strcat("s=",num2str(s)) );
imwrite(uint8(A),strcat(['exo6-' num2str(s) '.jpg'])
    ,'jpg');

[U,S,V]=svd(A);

% on fait des economies de stockage si "s" est < a "
    economie" :
economie=row*col/(row+col+1)
vsSignif=sum(sum( (S./S(1,1))>1.e-3 ))
```

```
n=2;
for s=[vsSignif,floor(economie),100,10]
   subplot(5,2,2*n-1)
    X=U(:,1:s)*S(1:s,1:s)*(V(:,1:s))';
    imshow(uint8(X));
    title ( strcat("s=",num2str(s)) );
    imwrite(uint8(X),strcat(['exo6-' num2str(s) '.
        jpg']),'jpg');
   subplot(5,2,2*n)
    erreur=abs(X-A);
    somerr=sum(erreur(:));
    m=min(erreur(:));
    M=max(erreur(:));
    erreur=255-255/(M-m)*(erreur-m);
    imshow(uint8(erreur));
    imwrite(uint8(erreur),strcat(['exo6-' num2str(s)
        'E.jpg']),'jpg');
   n+=1;
end
```

On a $\max\left\{ s \in [0;r] \;\middle|\; s < \frac{np}{n+p+1} \right\} = 255$: on fait des économies de stockage tant qu'on garde au plus les premières 255 valeurs singulières.
La photo de Lena, de taille 512×512, possède 275 valeurs singulières «significatives».

A.15.4. Les «mauvaises» propriétés des nombres flottants et la notion de précision

L'expérimentation numérique dans les sciences est un sujet passionnant et un outil fort utile, devenu indispensable pour certains scientifiques. Malgré la puissance vertigineuse de calcul de nos ordinateurs aujourd'hui, et encore plus de certains centres de calculs, on aurait tort d'oublier complètement la théorie et de trop se moquer de comment fonctionne la machine, au risque d'avoir quelques surprises…
L'expérimentation numérique dans les sciences est un sujet passionnant et un outil fort utile, devenu indispensable pour certains scientifiques. Malgré la puissance vertigineuse de calcul de nos ordinateurs aujourd'hui, et encore plus de certains centres de calculs, on aurait tort d'oublier complètement la théorie et de trop se moquer de comment fonctionne la machine, au risque d'avoir quelques surprises…
Observons des calculs quelque peu surprenants :

```
format long
0.1 + 0.1 + 0.1 - 0.3 % ans = 5.55111512312578e-17
```

Que s'est-il passé? Tout simplement, les calculs effectués ne sont pas exacts et sont entachés d'erreurs d'arrondis. En effet, tout nombre réel possède un développement décimal soit fini soit illimité. Parmi les nombres réels, on peut alors distinguer les rationnels (dont le développement décimal est soit fini soit illimité et périodique à partir d'un certain rang) des irrationnels (dont le développement décimal est illimité et non périodique). Il est aisé de concevoir qu'il n'est pas possible pour un ordinateur de représenter de manière exacte un développement décimal illimité, mais même la représentation des développements décimaux finis n'est pas toujours possible. En effet, un ordinateur stocke les nombres non pas en base 10 mais en base 2. Or, un nombre rationnel peut tout à fait posséder un développement décimal fini et un développement binaire illimité!

Erreurs d'arrondis

```
1 / 3 - 1 / 4 - 1 / 12 % ans = -1.38777878078145e-17
```

Non-commutativité

```
1 + 1e-16 - 1 % ans = 0
-1 + 1e-16 + 1 % ans = 1.11022302462516e-16
```

Représentation décimale inexacte Dans l'exemple ci-dessous 1.2 n'est pas représentable en machine. L'ordinateur utilise «le flottant représentable le plus proche de 1.2»

```
1.2 - 1 - 0.2 % ans = -5.55111512312578e-17
```

Conséquences

- ⋆ On ne peut pas espérer de résultat exact
- ⋆ La précision du calcul dépend de beaucoup d'éléments
- ⋆ En général, pour éviter les pertes de précision, on essayera autant que faire se peut d'éviter :
  - ⋆ de soustraire deux nombres très proches
  - ⋆ d'additionner ou de soustraire deux nombres d'ordres de grandeur très différents. Ainsi, pour calculer une somme de termes ayant des ordres de grandeur très différents (par exemple dans le calcul des sommes partielles d'une série), on appliquera le principe dit "de la photo de classe" : les petits devant, les grands derrière.
- ⋆ **tester** `x == flottant` **est presque toujours une erreur**, on utilisera plutôt `abs(x-flottant)<1.e-10` par exemple.
- ⋆ "Si on s'y prend bien", on perd $\approx 10^{-16}$ en précision relative à chaque calcul $\Rightarrow$ acceptable par rapport à la précision des données.
- ⋆ "Si on s'y prend mal", le résultat peut être complètement faux!

Illustrons ce problème d'arrondis en partant de l'identité suivante :

$$xy = \left(\frac{x+y}{2}\right)^2 - \left(\frac{x-y}{2}\right)^2.$$

Dans le programme suivant on compare les deux membres de cette égalité pour des nombres x et y de plus en plus grands :

```
prod=@(x,y)[x*y]
diff=@(x,y)[((x+y)/2)^2-((x-y)/2)^2]
a = 6553.99;
b = a+1;
A=zeros(6,3);
for i=1:6
   produit = prod(a,b);
   difference = diff(a,b);
   A(i,1)=a;
   A(i,2)=b;
   A(i,3)=produit-difference;
   b = a+1;
   a=produit;
end
printf("a | b | ab-((a+b)/2)^2-((a-b)/2)^2\n")
disp(A)
```

On constate que la divergence est spectaculaire.

Remarque

Voici deux exemples de désastres causés par une mauvaise gestion des erreurs d'arrondi :

- ⋆ Le 25 février 1991, pendant la Guerre du Golfe, une batterie américaine de missiles Patriot, à Dharan (Arabie Saoudite), a échoué dans l'interception d'un missile Scud irakien. Le Scud a frappé un baraquement de l'armée américaine et a tué 28 soldats. La commission d'enquête a conclu à un calcul incorrect du temps de parcours, dû à un problème d'arrondi. Les nombres étaient représentés en virgule fixe sur 24 bits. Le temps était compté par l'horloge interne du système en 1/10 de seconde. Malheureusement, 1/10 n'a pas d'écriture finie dans le système binaire : $1/10 = 0,1$ (dans le système décimal) $= 0,0001100110011001100110011\ldots$ (dans le système binaire). L'ordinateur de bord arrondissait 1/10 à 24 chiffres, d'où une petite erreur dans le décompte du temps pour chaque 1/10 de seconde. Au moment de l'attaque, la batterie de missile Patriot était allumée depuis environ 100 heures, ce qui avait entraîné une accumulation des erreurs d'arrondi de 0,34 s. Pendant ce temps, un missile Scud parcourt environ 500 m, ce qui explique que le Patriot soit passé à côté de sa cible.
- ⋆ Le 4 juin 1996, une fusée Ariane 5 a explosé 40 secondes après l'allumage. La fusée et son chargement avaient coûté 500 millions de dollars. La commission d'enquête a rendu son rapport au bout de deux semaines. Il s'agissait d'une erreur de programmation dans le système inertiel de référence. À un moment donné, un nombre codé en virgule flottante sur 64 bits (qui représentait la vitesse horizontale de la fusée par rapport à la plate-forme de tir) était converti en un entier sur 16 bits. Malheureusement, le nombre en question était plus grand que 32768, le plus grand entier que l'on peut coder sur 16 bits, et la conversion a été incorrecte.

☆ Exercice A.35
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 (~=) 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 : $1.1102 \times 10^{-16} = \epsilon_M/2$. Il existe donc au moins un nombre b différent de 0 tel que $a+b=a$.

☆ Exercice A.36
Calculer analytiquement et numériquement les premiers 100 termes des suites suivantes :

$$\begin{cases} u_0 = \frac{1}{4}, \\ u_{n+1} = 5u_n - 1, \end{cases} \qquad \begin{cases} v_0 = \frac{1}{5}, \\ v_{n+1} = 6v_n - 1, \end{cases} \qquad \begin{cases} w_0 = \frac{1}{3}, \\ w_{n+1} = 4v_w - 1. \end{cases}$$

Correction
Clairement $u_i = \frac{1}{4}$, $v_i = \frac{1}{5}$ et $w_i = \frac{1}{3}$ pour tout $i \in \mathbb{N}$. Cependant, lorsqu'on calcul les premiers 100 termes de ces deux suites avec Python (ou avec un autre langage de programmation) on a quelques surprises.
Si on écrit

```
n=30;
u=zeros(n+1,1);
u(1) = 1/4;
for i=1:n
    u(i+1) = 5*u(i)-1;
end
disp(u)
```

on trouve bien $u_i = 0.25$ pour tout $i = 0, \ldots$.
Mais si on écrit

```
n=30;
v=zeros(n+1,1);
v(1) = 1/5;
for i=1:n
    v(i+1) = 6*v(i)-1;
end
disp(v)
```

on obtient $v_i \simeq 0.2$ pour $i = 0, \ldots, 5$, ensuite les erreurs d'arrondis commencent à se voir.
De même

```
n=41;
w=zeros(n+1,1);
w(1) = 1/3;
for i=1:n
    w(i+1) = 4*w(i)-1;
end
disp(v)
```

À la vingtième répétition, le résultat est $w_{20} = 0.333328247070312$ ce qui est déjà assez éloigné de $1/3$. À la quarantième répétition de la ligne, le résultat est $w_{40} = -5592405$ ce qui n'a plus rien à voir. En fait, l'erreur sur l'arrondi se cumule et le résultat devient complètement absurde.

☆ Exercice A.37 (Suite de Muller)

Considérons la suite

$$\begin{cases} x_0 = 4, \\ x_1 = 4.25, \\ x_{n+1} = 108 - \frac{815 - \frac{1500}{x_{n-1}}}{x_n}. \end{cases}$$

On peut montrer que $\lim_{n\to+\infty} x_n = 5$ (voir par exemple https://scipython.com/blog/mullers-recurrence/)
Qu'obtient-on numériquement ?

Correction

```
x=[4, 4.25];
for i=3:30
    x(i)=108-(815-(1500/x(end-1)))/x(end);
end
disp(x)
```

☆ Exercice A.38 (Évaluer la fonction de Rump)

Évaluer au point $(x, y) = (77617, 33096)$ la fonction de deux variables suivante :

$$f(x,y) = \frac{1335}{4}y^6 + x^2(11x^2y^2 - y^6 - 121y^4 - 2) + \frac{11}{2}y^8 + \frac{x}{2y}$$

Correction

```
f = @(x,y) 1335*y^6/4+x^2*(11*x^2*y^2-y^6-121*y^4-2) + 11*y^8/2+x/(2*y) ;
f(77617,33096) % ans = -1.18059162071741e+21
```

Si on fait le calcul à la main ou on utilise le module de calcul formel, on trouve une valeur exacte d'environ -0.8273960599n.

☆ Exercice A.39 (Calcul d'intégrale par récurrence)

On veut approcher numériquement l'intégrale $I_n = \int_0^1 x^n e^{\alpha x} \mathrm{d}x$ pour $n = 50$. On remarque que, en intégrant par partie, on a

$$\int x^n e^{\alpha x} \mathrm{d}x = x^n \frac{1}{\alpha} e^{\alpha x} - \frac{n}{\alpha} \int x^{n-1} e^{\alpha x} \mathrm{d}x \tag{A.1}$$

ainsi

$$I_n = \int_0^1 x^n e^{\alpha x} \mathrm{d}x \tag{A.2}$$

$$= \frac{1}{\alpha} e^{\alpha} - \frac{n}{\alpha} I_{n-1} \tag{A.3}$$

On décide alors de calculer I_{50} par la suite récurrente suivante :

$$\begin{cases} I_0 = \frac{e^{\alpha}-1}{\alpha}, \\ I_{n+1} = \frac{1}{\alpha} e^{\alpha} - \frac{n+1}{\alpha} I_n, \quad \text{pour } n \in \mathbb{N}. \end{cases}$$

Écrire un programme pour calculer cette suite. Comparer le résultat numérique avec la limite exacte $I_n \to 0$ pour $n \to +\infty$.

Correction

Si on calcule I_n avec la formule de récurrence avec $\alpha = 1$, on remarque que $0 < I_{n+1} < I_n$ pour $n < 17$, mais $I_{18} < 0$ et la suite est instable. On a le même comportement pour les autres valeurs de α.

```
alpha=1;
myInt=[(exp(alpha)-1)/alpha]
for n=1:20
   myInt(n+1)=exp(alpha)/alpha - (n)*myInt(n)/alpha;
end
disp(myInt)
plot([0:20],myInt)
```

La suite obtenue avec le programme ci-dessus ne tend pas vers zéro quand n tend vers l'infini. Pourquoi un tel comportement numérique ? Ce comportement est une conséquence directe de la propagation des erreurs d'arrondi : en passant de I_n à I_{n+1}, l'erreur numérique (accumulation des erreurs de représentation et des premiers calculs) est multipliée par n :

$$\begin{aligned}\varepsilon_{n+1} &= I_{n+1}^{\text{exacte}} - I_{n+1}^{\text{approx}} \\ &= \left(\frac{1}{\alpha}e^{\alpha} - \frac{n+1}{\alpha} I_n^{\text{exacte}}\right) - \left(\frac{1}{\alpha}e^{\alpha} - \frac{n+1}{\alpha} I_n^{\text{approx}}\right) \\ &= -\frac{n+1}{\alpha}(I_n^{\text{exacte}} - I_n^{\text{approx}}) \\ &= -\frac{n+1}{\alpha}\varepsilon_n\end{aligned}$$

L'erreur numérique $|\varepsilon_n|$ sur l'évaluation de I_n croit donc comme $\frac{n!}{\alpha^n}|\varepsilon_0|$.

☆ Exercice A.40 (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

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

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

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

Dans Octave nous utiliserons le vecteur **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 `function` suivantes :

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 **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

   $$\begin{aligned} p(x) &= 1 + 2x + 3x^2 \\ \mathbf{p} &= \text{coord}(p, \mathscr{C}_2) = (1, 2, 3) \end{aligned} \qquad \begin{aligned} p(-1) &= 1 + 2 \times (-1) + 3 \times ((-1)^2) = 1 - 2 + 3 = 2 \\ p(0) &= 1 + 2 \times 0 + 3 \times (0^2) = 1 + 0 + 0 = 1 \\ p(1) &= 1 + 2 \times 1 + 3 \times (1^2) = 1 + 2 + 3 = 6 \\ p(2) &= 1 + 2 \times 2 + 3 \times (2^2) = 1 + 4 + 12 = 17 \end{aligned}$$

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 $\mathbf{s} = (2,0,3)$:

$$\begin{aligned} p(x) &= 1+2x+3x^2 & \mathbf{p} &= \text{coord}(p,\mathscr{C}_2) = (1,2,3) \\ q(x) &= 1-2x & \mathbf{q} &= \text{coord}(q,\mathscr{C}_1) = (1,-2) \implies \mathbf{q} = \text{coord}(q,\mathscr{C}_2) = (1,-2,0) \\ s(x) = p(x)+q(x) &= 2+3x^2 & \mathbf{s} &= \text{coord}(p+q,\mathscr{C}_2) = (2,0,3) \end{aligned}$$

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 $\mathbf{u} = (1,-2,3,-6)$.

$$\begin{aligned} p(x) &= 1+3x^2 & \mathbf{p} &= \text{coord}(p,\mathscr{C}_2) = (1,0,3) \\ q(x) &= 1-2x & \mathbf{q} &= \text{coord}(q,\mathscr{C}_2) = (1,-2,0) \\ u(x) = p(x)\times q(x) = 1\times p(x) - 2x\times p(x) &= 1-2x+3x^2-6x^3 & \mathbf{u} &= \text{coord}(p\times q,\mathscr{C}_3) = (1,-2,3,-6) \end{aligned}$$

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 $\mathbf{d} = (2,12)$.

$$\begin{aligned} p(x) &= 1+2x+6x^2 & \mathbf{p} &= \text{coord}(p,\mathscr{C}_2) = (1,2,6) \\ d(x) = p'(x) &= 2+12x & \mathbf{d} &= \text{coord}(d,\mathscr{C}_1) = (2,12) \end{aligned}$$

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 $\mathbf{v} = (0,1,1,2)$.

$$\begin{aligned} p(x) &= 1+2x+6x^2 & \mathbf{p} &= \text{coord}(p,\mathscr{C}_2) = (1,2,6) \\ v(x) = \int_0^x p(t)\,\mathrm{d}t = \int_0^x 1+2t+6t^2\,\mathrm{d}t &= x+x^2+2x^3 & \mathbf{v} &= \text{coord}(v,\mathscr{C}_3) = (0,1,1,2) \end{aligned}$$

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 $c = 18$:

$$\begin{aligned} p(x) &= 1+2x+6x^2 \\ c = \int_a^b p(t)\,\mathrm{d}t = \int_0^b p(t)\,\mathrm{d}t - \int_0^a p(t)\,\mathrm{d}t &= v(b)-v(a) = b+b^2+2b^3-a-a^2-2a^3 = 18. \end{aligned}$$

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`.

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

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

$$p(x) = \sum_{i=0}^{n} a_i x^i = \sum_{i=0}^{m} a_i x^i + \sum_{i=m+1}^{n} a_i x^i \qquad \text{coord}(p, \mathscr{C}) = (a_0, a_1, a_2, \dots, a_m, a_{m+1}, \dots, a_n)$$

$$q(x) = \sum_{i=0}^{m} b_i x^i = \sum_{i=0}^{m} b_i x^i + \sum_{i=m+1}^{n} 0 \times x^i \qquad \text{coord}(q, \mathscr{C}) = (b_0, b_1, b_2, \dots, b_m)$$

$$(p+q)(x) = \sum_{i=0}^{m} (a_i + b_i) x^i + \sum_{i=m+1}^{n} a_i x^i \qquad \text{coord}(p+q, \mathscr{C}) = (a_0 + b_0, a_1 + b_1, a_2 + b_2, \dots, a_m + b_m, a_{m+1}, \dots, a_n)$$

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

$$p(x) = \sum_{i=0}^{n} a_i x^i \qquad d(x) = p'(x) = \sum_{i=0}^{n} i\, a_i x^{i-1}$$

$$\text{coord}(p, \mathscr{C}_n) = (a_0, a_1, a_2, \dots, a_n) \qquad \text{coord}(d, \mathscr{C}_{n-1}) = (a_1, 2a_2, \dots, n a_n)$$

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

$$p(x) = \sum_{i=0}^{n} a_i x^i \qquad v(x) = \int_0^x p(t)\,\mathrm{d}t = \sum_{i=0}^{n} a_i \int_0^x t^i\,\mathrm{d}t = \sum_{i=0}^{n} a_i \frac{x^{i+1}}{i+1}$$

$$\text{coord}(p, \mathscr{C}_n) = (a_0, a_1, a_2, \dots, a_n) \qquad \text{coord}(v, \mathscr{C}_{n+1}) = \left(0, \frac{a_0}{0+1}, \frac{a_1}{1+1}, \frac{a_2}{2+1}, \dots, \frac{a_n}{n+1}\right)$$

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])
```