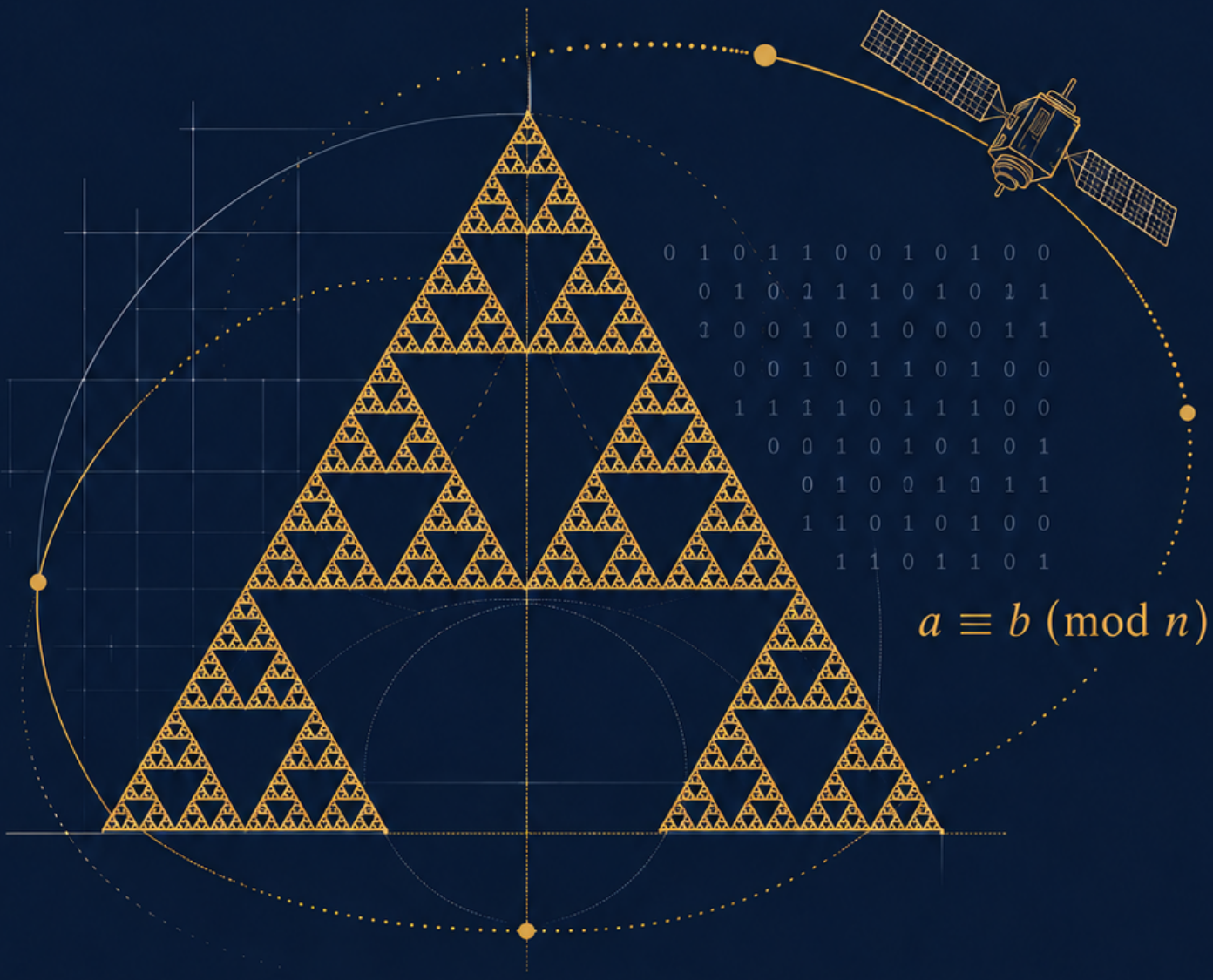
Mathématiques et technologie

Des notions du secondaire aux applications modernes



Jérôme Soucy

UNIVERSITÉ LAVAL

DÉPARTEMENT DE MATHÉMATIQUES ET DE STATISTIQUE

Au sujet de ces notes

Ces notes sont destinées aux étudiantes et étudiants du baccalauréat en enseignement secondaire de l'Université Laval qui suivent le cours *MAT-2901 : Mathématiques et technologie*. Dans cette perspective, nous nous intéresserons aussi aux liens entre les notions étudiées et les mathématiques enseignées dans les écoles secondaires québécoises.

Les notes de Jean-Jacques Gervais [Ger08], utilisées pendant plusieurs années dans ce cours, ont servi de point de départ à une partie de ce document. Le premier chapitre développe notamment le contenu de leur annexe 1, consacrée aux transformations affines du plan. Le troisième reprend des éléments de leur premier chapitre, qui porte sur la construction de fractales.

Licence CC BY 4.0

© 2026 Jérôme Soucy. Ce document est mis à disposition selon les termes de la licence Creative Commons Attribution 4.0 International (CC BY 4.0). Vous êtes libre de le copier, de le redistribuer, de le modifier et de l'adapter, pour tout usage, à condition de citer l'auteur, d'indiquer si des modifications ont été apportées et de fournir un lien vers la licence. Les programmes qui l'accompagnent (code Python, filtres et application interactive) sont distribués sous la licence MIT. Les portraits et les jeux de données conservent leur licence d'origine ; leur source est indiquée dans leur légende ou à leur première utilisation.

Pour citer ce document : Jérôme Soucy, *Mathématiques et technologie : des notions du secondaire aux applications modernes*, 2026.

Table des matières

1 Les transformations du plan **8**
- 1.1 Introduction . . . 8
- 1.2 Qu'est-ce qu'une transformation du plan ? . . . 9
  - 1.2.1 Exemples et contre-exemples . . . 11
  - 1.2.2 Composer des transformations . . . 12
- 1.3 Le groupe des transformations du plan . . . 13
- 1.4 Du plan géométrique au plan des coordonnées . . . 14
- 1.5 Les transformations affines . . . 15
  - 1.5.1 Ce qu'une transformation affine conserve . . . 16
  - 1.5.2 Visualiser une transformation affine : l'image du carré unité . . . 17
- 1.6 Un catalogue d'applications affines . . . 20
  - 1.6.1 Translations . . . 20
  - 1.6.2 Homothéties . . . 20
  - 1.6.3 Rotations . . . 22
  - 1.6.4 Réflexions . . . 23
  - 1.6.5 Cisaillements . . . 24
  - 1.6.6 Contractions et dilatations non uniformes . . . 24
  - 1.6.7 Un catalogue en images . . . 24
  - 1.6.8 Le cas non inversible . . . 24
  - 1.6.9 Le théorème de décomposition . . . 26
- 1.7 Les coordonnées homogènes et leurs applications en infographie . . . 27
- 1.8 Isométries et similitudes . . . 29
  - 1.8.1 Les propriétés des similitudes . . . 30
  - 1.8.2 Classification des isométries du plan . . . 31
  - 1.8.3 La décomposition en réflexions . . . 31
- 1.9 Déterminer une transformation à partir de points . . . 33
- 1.10 Les transformations dans l'enseignement secondaire québécois . . . 35
  - 1.10.1 Où sont-elles dans le programme ? . . . 35
  - 1.10.2 Traduire la notation du secondaire . . . 36
  - 1.10.3 Les apports du point de vue matriciel . . . 36
  - 1.10.4 Les difficultés d'apprentissage, et d'où elles viennent . . . 37
  - 1.10.5 Quelques activités pour la classe . . . 38

2 Espaces métriques **42**
- 2.1 Qu'est-ce qu'une distance ? . . . 43
  - 2.1.1 Les distances euclidienne et de Manhattan . . . 44
  - 2.1.2 Une métrique appliquée : la distance sur la sphère . . . 45
  - 2.1.3 Une métrique appliquée : mesurer les erreurs . . . 45
- 2.2 Ouverts, fermés et compacts . . . 47
  - 2.2.1 Boules et ensembles ouverts . . . 47
  - 2.2.2 Ensembles fermés . . . 48

2.2.3 Ensembles compacts . . . 49
2.3 La distance de Hausdorff . . . 49
2.3.1 Le ε-voisinage et la somme de Minkowski . . . 49
2.3.2 Définition de la distance de Hausdorff . . . 50
2.4 Suites de Cauchy et complétude . . . 52
2.5 La notion de point fixe . . . 55
2.6 Contractions . . . 57
2.7 Le théorème du point fixe de Banach . . . 58
2.7.1 La méthode de Newton-Raphson . . . 62

3 Fractales et systèmes itératifs de fonctions 67
3.1 Des monstres mathématiques . . . 68
3.1.1 La courbe de von Koch . . . 68
3.1.2 Le triangle et le tapis de Sierpiński . . . 70
3.2 Autosimilarité . . . 72
3.3 La dimension de similitude . . . 75
3.3.1 Compter les copies . . . 76
3.3.2 Ce que mesure la dimension . . . 77
3.4 Systèmes itératifs de fonctions . . . 79
3.4.1 L'opérateur de Hutchinson est une contraction . . . 79
3.4.2 Le théorème de Hutchinson . . . 80
3.4.3 Trouver un SIF : le théorème du collage . . . 82
3.5 La dimension de similitude généralisée . . . 85
3.6 Dessiner un attracteur à l'ordinateur . . . 87
3.6.1 L'algorithme déterministe . . . 88
3.6.2 Le jeu du chaos . . . 89
3.6.3 Choisir les probabilités . . . 91
3.7 Des monstres aux modèles . . . 92
3.7.1 La fonction de Weierstrass . . . 92
3.7.2 Benoît Mandelbrot et la géométrie de la nature . . . 94
3.7.3 Les fractales et la finance . . . 95

4 Les chaînes de Markov et l'algorithme PageRank de Google 101
4.1 Rappels sur les probabilités . . . 102
4.1.1 Probabilités conditionnelles . . . 102
4.1.2 Variables aléatoires discrètes et espérance . . . 103
4.2 Un premier modèle : la météo à deux états . . . 105
4.2.1 Prévoir le temps des jours suivants . . . 105
4.2.2 L'état d'équilibre . . . 106
4.2.3 Le calcul avec Python . . . 107
4.3 Les chaînes de Markov . . . 108
4.3.1 Définition . . . 108
4.3.2 Ville, banlieue et campagne . . . 110
4.3.3 Et si l'avenir dépendait de plusieurs jours ? . . . 111
4.4 Matrices stochastiques et états d'équilibre . . . 112
4.4.1 La valeur propre 1 . . . 113
4.4.2 Deux chaînes qui ne convergent pas . . . 113
4.4.3 Le théorème de convergence . . . 114
4.5 L'algorithme PageRank . . . 115
4.5.1 L'internaute aléatoire . . . 116
4.5.2 Les pages sans liens sortants . . . 118
4.5.3 Les régions piégeantes et la téléportation . . . 119

4.5.4 La matrice Google . . . 121
4.6 Calculer le PageRank : la méthode des puissances . . . 121
4.6.1 Pourquoi la méthode converge : retour au théorème de Banach . 122

5 L'analyse en composantes principales 129
5.1 Un peu de statistique descriptive . . . 129
5.1.1 Individus, variables et tableau de données . . . 129
5.1.2 Moyenne, variance et écart type . . . 130
5.1.3 Covariance et corrélation . . . 131
5.2 La droite qui représente le mieux un nuage de points . . . 134
5.2.1 Deux façons de mesurer l'écart à une droite . . . 134
5.2.2 On peut se ramener à des données centrées . . . 135
5.2.3 Projeter sur une droite : le théorème de Pythagore . . . 135
5.2.4 La variance dans une direction quelconque . . . 137
5.2.5 Le lien avec les vecteurs propres . . . 137
5.3 Des données dans l'espace . . . 140
5.3.1 Les manchots de l'archipel Palmer . . . 140
5.3.2 Regarder le nuage sous différents angles . . . 141
5.3.3 Projeter sur un plan . . . 142
5.3.4 Un problème d'unités : réduire les variables . . . 143
5.3.5 Le meilleur plan . . . 143
5.4 L'analyse en composantes principales . . . 145
5.4.1 Le tableau centré et la matrice de covariance . . . 145
5.4.2 Axes principaux et composantes principales . . . 146
5.4.3 Combien d'axes conserver ? . . . 148
5.4.4 Le calcul avec Python . . . 149
5.4.5 Un peu d'histoire . . . 150
5.5 Au-delà de trois dimensions . . . 151
5.5.1 Le décathlon . . . 151
5.5.2 Lire le plan principal . . . 153
5.5.3 D'autres applications . . . 154

6 Numération et stockage de l'information 159
6.1 Écrire un nombre dans une base quelconque . . . 159
6.1.1 La numération en base dix . . . 159
6.1.2 Une base quelconque . . . 159
6.2 Le binaire, langue des circuits . . . 162
6.2.1 Pourquoi deux symboles ? . . . 162
6.2.2 Calculer en binaire . . . 164
6.2.3 Des entiers de taille fixe . . . 164
6.2.4 Et les nombres à virgule ? . . . 164
6.2.5 La norme IEEE 754 . . . 165
6.3 L'octal et l'hexadécimal, abréviations du binaire . . . 166
6.3.1 Le problème de la lisibilité . . . 166
6.3.2 L'octal . . . 167
6.3.3 L'hexadécimal . . . 167
6.3.4 Exemple phare : les codes de couleur du Web . . . 168
6.4 Encoder du texte . . . 170
6.4.1 Un encodage, c'est une convention . . . 170
6.4.2 Du télégraphe à l'ASCII . . . 170
6.4.3 Les accents : un octet pour tout le monde ? . . . 171
6.4.4 Unicode : un numéro pour chaque caractère . . . 171

6.4.5 L'UTF-8 . . . 171
6.5 Mesurer l'information . . . 173
6.5.1 Le bit et l'octet . . . 173
6.5.2 Préfixes décimaux et préfixes binaires . . . 174
6.5.3 Où sont passés les gigaoctets ? . . . 174
6.6 Mémoire vive et mémoire de stockage . . . 175
6.6.1 Le bureau et la bibliothèque . . . 175
6.6.2 Comment un bit est-il inscrit ? . . . 176
6.7 Un historique des capacités de stockage . . . 178
6.8 Les images numériques . . . 179
6.8.1 Les images matricielles . . . 179
6.8.2 La compression . . . 181
6.8.3 Les images vectorielles . . . 182

7 La cryptographie 187
7.1 Le vocabulaire de la cryptologie . . . 187
7.1.1 Chiffrer, déchiffrer, décrypter . . . 187
7.1.2 Le principe de Kerckhoffs . . . 188
7.1.3 Des lettres aux nombres . . . 189
7.2 Les chiffrements par décalage . . . 189
7.2.1 Le chiffre de César . . . 189
7.2.2 Une attaque exhaustive . . . 190
7.3 Le chiffre affine . . . 191
7.3.1 Chiffrer avec une fonction affine . . . 191
7.3.2 Quelles clés sont admissibles ? . . . 192
7.3.3 Une attaque à texte clair connu . . . 193
7.3.4 Activité : retrouver la clé . . . 193
7.4 La substitution et l'analyse des fréquences . . . 194
7.4.1 Un nombre astronomique de clés . . . 194
7.4.2 L'analyse des fréquences . . . 194
7.4.3 Casser un chiffre affine par les fréquences . . . 195
7.5 Le chiffre de Vigenère . . . 197
7.5.1 Plusieurs décalages . . . 197
7.5.2 Trouver la longueur de la clé . . . 197
7.5.3 Le masque jetable . . . 199
7.6 Le chiffre de Hill . . . 199
7.6.1 Chiffrer par blocs avec une matrice . . . 199
7.6.2 Inverser une matrice modulo 26 . . . 200
7.6.3 Une attaque à texte clair connu . . . 202
7.7 Enigma, la mécanisation du chiffrement . . . 203
7.7.1 La machine . . . 203
7.7.2 Une involution sans point fixe . . . 204
7.7.3 Le nombre de clés . . . 205
7.8 La cryptographie à clé publique . . . 206
7.8.1 Le problème de l'échange des clés . . . 206
7.8.2 L'échange de clés de Diffie et Hellman . . . 207
7.8.3 Le chiffrement RSA . . . 207
7.8.4 L'exponentiation rapide . . . 209
7.8.5 Pourquoi RSA est-il sûr ? . . . 210
7.8.6 La signature numérique . . . 210

Annexes **216**

partie Calculer les transformations avec Python **216**
Représenter une figure et une transformation . . . 216
Une petite bibliothèque de transformations . . . 217
Composer : les coordonnées homogènes en pratique . . . 217
Résoudre le problème des deux triangles . . . 218
Des réponses exactes avec `sympy` . . . 218
Reconnaître une similitude . . . 219
Dessiner . . . 219
Reconnaître le genre et l'angle d'une similitude . . . 220

partie La rédaction du chapitre 2 avec l'assistance d'une intelligence artificielle **222**
Les documents de départ . . . 222
Le mandat initial . . . 222
Valider la compréhension avant d'écrire . . . 223
Une relecture détaillée . . . 223
Raffiner les figures . . . 224
Restructurer le contenu . . . 224
Les détails techniques . . . 225
En rétrospective . . . 225

partie Compléments théoriques sur les SIF **226**
Démonstration du théorème 3.1 . . . 226
Démonstration du théorème 3.3 . . . 227

partie Rappels d'algèbre linéaire **228**
partie Vecteurs de $\mathbb{R}^n$ . . . 228
partie Combinaisons linéaires et espace engendré . . . 229
partie Produit scalaire, norme et angle . . . 231
partie Matrices et opérations matricielles . . . 232
partie Le produit d'une matrice par un vecteur . . . 232
partie Le produit de deux matrices . . . 232
partie Systèmes linéaires et méthode de Gauss–Jordan . . . 233
partie Ce que la géométrie nous apprend . . . 234
partie La méthode de Gauss–Jordan . . . 234
partie Indépendance linéaire, bases et dimension . . . 237
partie Noyau et image . . . 238
partie Le déterminant . . . 239
partie L'inverse d'une matrice . . . 241
partie Valeurs propres et vecteurs propres . . . 243
partie Le polynôme caractéristique . . . 244
partie Diagonalisation . . . 245
partie Le cas des matrices symétriques . . . 246
partie Diagonaliser avec Python . . . 247

partie Éléments de théorie des nombres **251**
partie Divisibilité et division euclidienne . . . 251
partie La division euclidienne . . . 252
partie Diviseurs communs et multiples communs . . . 253
partie Nombres premiers . . . 254

partie Une infinité de nombres premiers 255
partie Existence de la factorisation première 256
partie Le PGCD et l'algorithme d'Euclide 257
partie Une interprétation géométrique : paver un rectangle . 259
partie Le théorème de Bachet–Bézout 260
partie Le lemme de Gauss et l'unicité de la factorisation 262
partie Le théorème fondamental de l'arithmétique 263
partie Les équations diophantiennes linéaires 264
partie Les congruences 265
partie Inverses modulaires et congruences linéaires 267
partie Le petit théorème de Fermat 270
partie L'indicatrice d'Euler et le théorème d'Euler 271
partie Le théorème des restes chinois 274

Solutions et réponses **280**

Bibliographie **325**

1 Les transformations du plan

1.1 Introduction

Une transformation du plan déplace ou déforme les figures selon une règle précise. Certaines, comme les rotations et les réflexions, préservent les distances ; d'autres, comme les transformations affines, préservent plutôt l'alignement et le parallélisme. En les représentant par des fonctions et des matrices, on peut étudier ces transformations dans un cadre commun et comprendre quelles propriétés elles conservent.

Ce point de vue intervient notamment dans la construction de fractales. Par exemple, le triangle de Sierpinski peut être reconstruit à partir de trois transformations affines du plan, décrites par dix-huit nombres. La figure obtenue est un ensemble invariant associé à ces transformations [Bar93]. Représenter une image par un petit nombre de transformations est aussi à la base de la *compression fractale d'images* [Kom97].

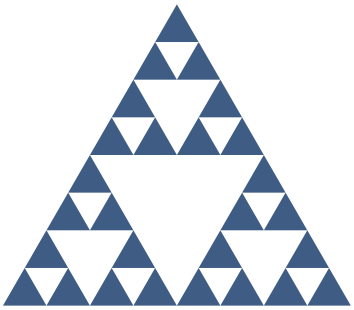


FIGURE 1.1 : Le triangle de Sierpinski peut être reconstruit à partir de trois transformations affines du plan, décrites par dix-huit nombres. Cette représentation illustre le principe de la compression fractale.

L'étude des transformations présente aussi un intérêt pour l'enseignement. On les rencontre à plusieurs moments du parcours scolaire québécois : la réflexion apparaît au troisième cycle du primaire, la translation et la rotation en première secondaire, l'homothétie et les figures semblables en deuxième et troisième secondaire, et le langage vectoriel qui les unifie en cinquième secondaire [Min16][1]. Ces transformations sont généralement présentées séparément, avec des constructions et un vocabulaire propres à chacune : le pliage ou le miroir pour la réflexion, le calque et le compas pour la rotation, le quadrillage pour l'homothétie. Cette approche permet de développer l'intuition avant d'introduire un formalisme plus général. Cependant, elle ne fait pas nécessairement ressortir les propriétés communes aux transformations étudiées. Sans proposer de présenter une telle synthèse aux élèves du secondaire, il est utile que leurs enseignants puissent la faire. Nous chercherons

1. Disponible en ligne : https://cdn-contenu.quebec.ca/cdn-contenu/education/pfeq/secondaire/progressions-apprentissages/PFEQ-progression-apprentissages-mathematique-secondaire.pdf

donc à dégager un cadre commun qui permette de mieux comprendre ces transformations et leurs propriétés.

Note historique 1.1 Le programme d'Erlangen

En 1872, Felix Klein (1849–1925) prend possession de sa chaire à l'université d'Erlangen. Il y propose une façon de classer les géométries qui aura une influence considérable [Kle93][a]. Alors qu'on distinguait les géométries par leurs axiomes, Klein propose de les caractériser par leurs *transformations*. Dans cette perspective, une géométrie consiste à étudier les propriétés des figures qui sont conservées par les transformations d'un groupe donné.

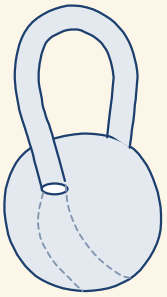


Felix Klein (1849–1925). Domaine public, *Wikimedia Commons*.

À titre d'exemple, les isométries préservent les longueurs, les angles et les aires. On retrouve ainsi la géométrie euclidienne étudiée au secondaire. Si on considère plutôt les similitudes, les longueurs peuvent changer, mais leurs *rapports* et les angles sont conservés : c'est le contexte des figures semblables. Les transformations affines forment un groupe plus vaste encore, pour lequel on conserve le parallélisme, l'alignement et les rapports de longueurs *sur une même droite*. Enfin, l'étude des propriétés conservées par les déformations continues, comme la présence d'un trou, relève de la topologie.

Ainsi, lorsqu'on élargit le groupe de transformations considéré, on restreint les propriétés géométriques qui sont conservées. Nous organiserons l'étude des transformations suivant cette hiérarchie. À la section 1.10, nous verrons aussi comment elle permet de situer les transformations introduites dans le programme québécois.

Le nom de Klein est également associé à un objet plus connu du grand public[b], la **bouteille de Klein**, qu'il décrit en 1882. Cette surface fermée, qui n'a ni intérieur ni extérieur, rappelle le ruban de Möbius. Pour la représenter sans qu'elle se recoupe, il faut se placer dans un espace à quatre dimensions. Les modèles que l'on fabrique ou que l'on dessine présentent donc une intersection, comme c'est le cas ici. Selon une anecdote souvent rapportée, son nom proviendrait d'une confusion entre *Fläche* (surface) et *Flasche* (bouteille).



a. Le texte original, en allemand, est disponible en ligne, en édition intégrale et libre de droits : https://www.gutenberg.org/ebooks/38033.

b. Pas si grand que ça quand même !

1.2 Qu'est-ce qu'une transformation du plan ?

Dans ce chapitre, nous travaillerons principalement dans le **plan euclidien**, noté Π ou $\mathbb{R}^2$. Il s'agit du plan usuel de la géométrie, dans lequel on dispose d'une distance $d(P, Q)$ entre deux points.

Définition 1.1 **Transformation géométrique du plan**

Une **transformation géométrique** du plan est une bijection du plan dans lui-même, c'est-à-dire une fonction bijective $f : \Pi \longrightarrow \Pi$. On note $\mathcal{T}$ l'ensemble de toutes les transformations du plan.

La définition fait intervenir deux notions qu'il convient de préciser : f doit être une *fonction* et cette fonction doit être *bijective*.

La notation $f : \Pi \longrightarrow \Pi$ désigne une fonction de Π vers Π. Cela signifie

(i) que f soit définie en chaque point $P \in \Pi$ (partout définie) ;
(ii) et qu'elle lui associe un unique point $f(P)$ (univoque).

Ces deux conditions expriment que chaque point de l'ensemble de départ possède une et une seule image. Pour que la fonction soit bijective, nous devons ajouter les deux conditions suivantes :

(iii) si $P \neq Q$, alors $f(P) \neq f(Q)$ (injectivité) ;
(iv) et pour tout $Q \in \Pi$, il existe $P \in \Pi$ tel que $Q = f(P)$ (surjectivité).

La première de ces conditions signifie que deux points distincts ont des images distinctes ; la seconde, que chaque point du plan d'arrivée possède un antécédent. Ainsi, dans la définition 1.1, une transformation est une fonction $f : \Pi \to \Pi$ qui satisfait les conditions (i) et (ii), ainsi que les conditions de bijectivité (iii) et (iv).

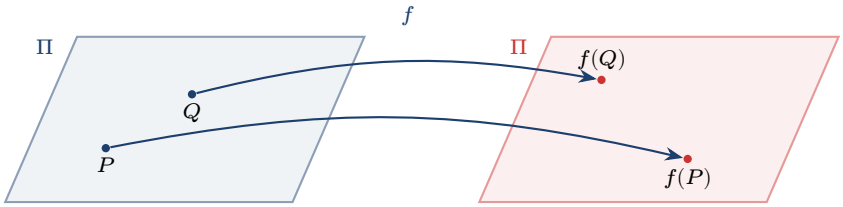


FIGURE 1.2 : Chaque point du plan possède une unique image par la transformation. Deux points distincts ont des images distinctes (injectivité) et chaque point du plan d'arrivée possède un antécédent (surjectivité).

Pour la classe 1.1 **Quelques précisions de vocabulaire**

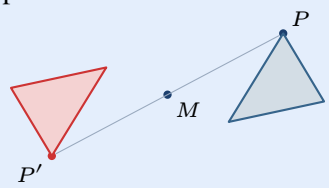
1. Lorsque $Q = f(P)$, on dit que Q est l'**image** de P par f, et que P est un **antécédent** de Q.
2. On dira souvent, dans un contexte géométrique, que « P s'est déplacé en Q ». Cette expression suggère un mouvement, alors que la définition porte sur une *correspondance* entre des points. Nous verrons à la section 1.10 que cette confusion entre « transformation » et « mouvement » est l'une des difficultés les mieux documentées chez les élèves — et chez les futurs enseignants [Yan11].
3. Une transformation agit sur *tout* le plan, pas seulement sur la figure qu'on a dessinée. Cette distinction peut être masquée par le vocabulaire employé en classe. Lorsqu'on dit qu'on « fait tourner un triangle », la transformation s'applique en fait au plan entier. Le triangle sert à représenter l'effet de cette transformation.

1.2.1 Exemples et contre-exemples

Pour mieux comprendre le rôle de chacune des conditions de la définition, examinons un exemple et deux contre-exemples.

Exemple 1.1 **Une symétrie centrale**

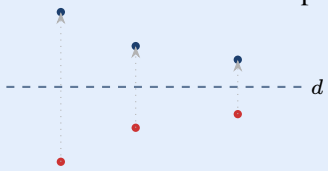
Soit M un point fixe et soit $f_1 : \Pi \to \Pi$ la correspondance qui envoie P sur l'unique point P' tel que M soit le milieu du segment $[PP']$; autrement dit, P' est le symétrique de P par rapport à M.



Vérifions que f_1 est une transformation du plan : chaque point P a exactement une image P', l'unique point tel que $\overrightarrow{MP'} = -\overrightarrow{MP}$, et chaque point est l'image d'exactement un point (en appliquant de nouveau la correspondance à l'image, on retrouve le point de départ). C'est la **symétrie centrale de centre** M. Elle préserve les longueurs, les angles et la forme des figures ; nous la retrouverons à la section 1.6 comme cas particulier d'une homothétie de rapport négatif.

Exemple 1.2 **Un repliage du plan**

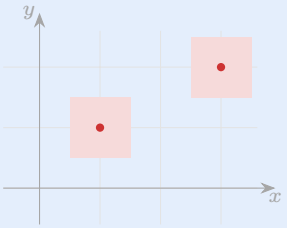
Soit d une droite fixe. Soit $f_2 : \Pi \to \Pi$ qui laisse fixe tout point situé du côté « nord » de d (droite comprise), et qui envoie tout point situé de l'autre côté sur son réfléchi par rapport à d : on obtient le même effet en pliant une feuille le long de d.



Cette correspondance définit une fonction, puisque chaque point possède une unique image. Elle n'est cependant pas injective : un point Q situé strictement au nord de d a deux antécédents, lui-même et son réfléchi au sud. Elle n'est pas surjective non plus : aucun point situé strictement au sud de d n'est jamais atteint, puisque les images sont toutes au nord ou sur la droite. La fonction ne satisfait donc pas la définition d'une transformation du plan.

Exemple 1.3 **L'arrondi au pixel près**

Soit $f_3 : \Pi \to \Pi$ qui envoie (x, y) sur le point de coordonnées entières le plus proche, c'est-à-dire sur $(\lfloor x \rceil, \lfloor y \rceil)$ où $\lfloor \cdot \rceil$ désigne l'arrondi à l'entier le plus proche. Cette correspondance illustre l'arrondi associé à l'affichage d'une image sur une grille de pixels.



Les points d'un carré de côté 1 centré en un point de coordonnées entières ont la même image. Ainsi, f_3 n'est pas injective. D'autre part, les points dont les coordonnées ne sont pas entières n'ont aucun antécédent, de sorte que f_3 n'est pas surjective. Il ne s'agit donc pas d'une transformation du plan.

1.2.2 Composer des transformations

Définition 1.2 **Égalité et composition**

Pour $f, g \in \mathcal{T}$, on a $f = g$ si et seulement si $f(P) = g(P)$ pour tout point $P \in \Pi$. On définit la **composée** $f \circ g : \Pi \to \Pi$ par la règle

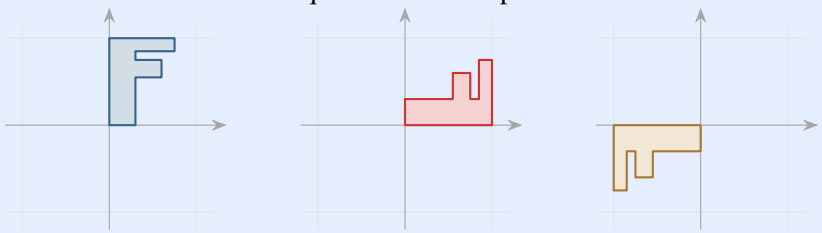
$$(f \circ g)(P) = f(g(P)).$$

Il faut porter attention à l'ordre de composition : dans $f \circ g$, on applique d'abord g. Cette convention découle de la notation fonctionnelle. Nous y reviendrons, car elle occasionne aussi des difficultés dans l'apprentissage des transformations au secondaire.

La composée de deux transformations du plan est encore une transformation du plan. On peut le vérifier en reprenant les quatre conditions de la définition 1.1 ; cette vérification est proposée à l'exercice 1.3.

Exemple 1.4 **La composition n'est pas commutative**

Considérons r la rotation de 90° autour de l'origine et R la réflexion par rapport à l'axe des x. En appliquant les deux composées au même motif, on obtient les figures ci-dessous. On constate que le résultat dépend de l'ordre choisi.



Les deux images sont symétriques l'une de l'autre par rapport à l'origine : ici $r \circ R \neq R \circ r$, et plus précisément $R \circ r = r^{-1} \circ R$.

1.3 Le groupe des transformations du plan

Les observations précédentes font ressortir trois propriétés de $\mathcal{T}$: la composition de deux de ses éléments appartient encore à cet ensemble, l'identité en fait partie et chaque transformation possède une réciproque. Précisons ces propriétés.

Théorème 1.1 **Propriétés de $(\mathcal{T}, \circ)$**

Soient $f, g, h \in \mathcal{T}$. Alors :

(i) **fermeture** : $f \circ g \in \mathcal{T}$;

(ii) **associativité** : $(f \circ g) \circ h = f \circ (g \circ h)$;

(iii) **identité** : $\mathrm{id}(P) = P$ appartient à $\mathcal{T}$ et vérifie $f \circ \mathrm{id} = \mathrm{id} \circ f = f$;

(iv) **inverse** : il existe $f^{-1} \in \mathcal{T}$ telle que $f \circ f^{-1} = f^{-1} \circ f = \mathrm{id}$.

En revanche, l'ordre compte : en général $f \circ g \neq g \circ f$.

Démonstration. Nous avons déjà établi la fermeture. Pour l'associativité, on utilise la propriété correspondante de la composition des fonctions : les deux membres évalués en P donnent $f(g(h(P)))$. Enfin, toute bijection admet une bijection réciproque : la condition de bijectivité assure donc l'existence de l'inverse. L'exemple 1.4 montre toutefois que la composition n'est pas commutative. □

Remarque 1.1 **Inverser une composée**

La réciproque d'une composée s'obtient en inversant l'ordre des fonctions :

$$(f \circ g)^{-1} = g^{-1} \circ f^{-1},$$

ce qui diffère en général de $f^{-1} \circ g^{-1}$. Une analogie familière permet de retenir cet ordre : après avoir mis ses bas, puis ses souliers, il faut enlever les souliers avant les bas pour revenir à la situation de départ.

Remarque 1.2 **Le mot groupe**

Les propriétés de fermeture et d'associativité, ainsi que l'existence d'un élément neutre et d'un inverse pour chaque élément, définissent la structure de **groupe**. Cette notion permet d'étudier dans un même cadre des objets aussi différents que des transformations, des nombres ou des permutations. L'étude de cette structure dépasse le cadre de ce cours : les groupes, leurs sous-groupes, les groupes de symétrie d'une figure et leurs applications aux pavages du plan sont l'objet du cours *MAT-2903 : Thèmes géométriques*. Nous en utiliserons cependant le vocabulaire : dans les sections qui suivent, nous dirons d'un ensemble de transformations qu'il est un **sous-groupe** de $\mathcal{T}$ lorsqu'il est lui-même stable par composition et par inversion.

Nous étudierons plusieurs familles de transformations, chacune étant incluse dans la précédente. La figure 1.3 présente leurs relations et servira de référence pour la suite du chapitre. Nous ferons aussi des liens avec les transformations étudiées au secondaire.

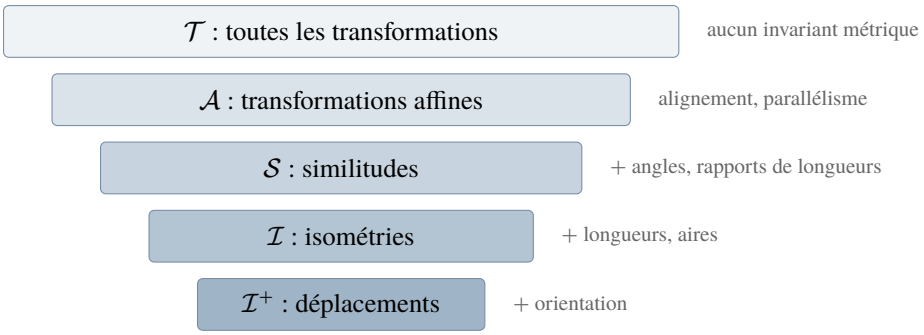


FIGURE 1.3 : Les principaux groupes de transformations du plan sont ordonnés par inclusion. En descendant dans le diagramme, on restreint les transformations permises et on conserve davantage de propriétés géométriques, suivant le point de vue proposé par Klein.

Remarque 1.3 **Où sont les applications affines non inversibles ?**

Chaque famille de la figure 1.3 est emboîtée dans $\mathcal{T}$, l'ensemble de *toutes* les transformations du plan ; or la définition 1.1 exige qu'une transformation soit bijective. Ainsi, chacune des familles représentées, y compris le groupe $\mathcal{A}$ que nous définirons plus loin, est constituée de bijections.
Il existe pourtant, comme nous le verrons à la section 1.6.8, des *applications* affines qui ne sont pas bijectives — la projection en est l'exemple type. Celles-ci ne préservent en général ni l'alignement ni le parallélisme (la preuve de la proposition 1.3 utilise explicitement l'inversibilité), et elles ne figurent dans *aucune* des familles de la figure 1.3 : n'étant pas bijectives, elles ne sont même pas dans $\mathcal{T}$. Il faut donc distinguer les deux expressions : « affine », sans autre précision, n'implique pas l'inversibilité tandis que c'est le cas pour « transformation affine ».

1.4 Du plan géométrique au plan des coordonnées

Jusqu'à présent, nous avons adopté un point de vue *synthétique* : les transformations étaient décrites géométriquement, sans faire appel aux coordonnées. C'est l'approche d'Euclide, que l'on retrouve aussi en première secondaire. Pour passer au point de vue *analytique*, nous allons choisir un repère. Les points seront alors représentés par des couples de nombres, et les transformations par des formules.

Choisissons une origine O et une base orthonormée $\{\mathbf{e}_1, \mathbf{e}_2\}$. Un point P du plan s'identifie alors au vecteur colonne de ses coordonnées, et nous écrirons indifféremment

$$P = (x, y) \qquad \text{ou} \qquad P = \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2 .$$

La distance entre deux points s'exprime par la norme euclidienne :

$$d(P, Q) = \|P - Q\| = \sqrt{(x_P - x_Q)^2 + (y_P - y_Q)^2}.$$

Ce passage à la géométrie analytique occupe une place importante en quatrième et cinquième secondaire. Une fois le repère choisi, on peut représenter un point par un couple de coordonnées, une droite par une équation et une région par des inéquations. Dans le cas des transformations que nous étudierons, nous utiliserons des matrices. Cette correspondance

entre géométrie et algèbre, associée aux travaux de Descartes, permet d'effectuer des calculs sur les transformations et de les programmer.

Remarque 1.4 **Points ou vecteurs ?**

En algèbre linéaire et en géométrie vectorielle, on distingue un *point* d'un *vecteur*. Nous utiliserons ici une convention qui simplifie les écritures : un point P et le vecteur $\overrightarrow{OP}$ sont représentés par la même colonne de nombres. Cet abus de notation est courant en infographie. Il ne faut toutefois pas perdre de vue la distinction entre les deux objets, notamment lorsqu'on cherche à comprendre pourquoi une translation n'est pas linéaire.

1.5 Les transformations affines

Nous allons maintenant étudier une famille de transformations du plan qui comprend toutes celles du programme du secondaire. Chacune de ces transformations peut être décrite à l'aide de six nombres.

Définition 1.3 **Application affine du plan**

Une **application affine** du plan est une fonction $w : \mathbb{R}^2 \to \mathbb{R}^2$ qui, dans la base standard, s'écrit sous la forme

$$w\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix}, \tag{1.1}$$

où $a, b, c, d, e, f \in \mathbb{R}$. On dit que (1.1) est la **représentation analytique** de w. En notation compacte, en posant $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ et $\mathbf{t} = \begin{pmatrix} e \\ f \end{pmatrix}$,

$$w\begin{pmatrix} x \\ y \end{pmatrix} = A\begin{pmatrix} x \\ y \end{pmatrix} + \mathbf{t}.$$

La matrice A s'appelle la **partie linéaire** de w, et le vecteur $\mathbf{t}$ sa **partie translation**.

Une application affine est donc la composée d'*une application linéaire suivie d'une translation*. Nous verrons que les translations, les rotations, les réflexions, les homothéties et les cisaillements peuvent tous s'écrire sous cette forme, de même que leurs composées.

Remarque 1.5 **Pourquoi « affine » et non « linéaire » ?**

Une application linéaire T vérifie $T(\mathbf{0}) = \mathbf{0}$: elle fixe nécessairement l'origine. Or une translation ne fixe aucun point. Les translations, qui sont pourtant les plus simples des transformations géométriques, ne sont donc *pas* linéaires. C'est précisément pour les inclure qu'on ajoute le terme constant $\mathbf{t}$, et qu'on parle d'applications affines. Nous verrons à la section 1.7 comment une dimension supplémentaire permet de représenter ces applications par des applications linéaires.

Rappelons de l'algèbre linéaire que le **noyau** d'une application linéaire $T : \mathbb{R}^2 \to \mathbb{R}^2$ est l'ensemble $\ker T = \{\mathbf{u} \in \mathbb{R}^2 : T(\mathbf{u}) = \mathbf{0}\}$; une application linéaire est injective

si et seulement si son noyau se réduit à $\{\mathbf{0}\}$, c'est- à-dire si et seulement si la matrice qui la représente est inversible. Nous utiliserons cette caractérisation pour démontrer la proposition suivante.

Proposition 1.2 **Inversibilité**

Une application affine de partie linéaire A et de partie translation $\mathbf{t}$ est bijective si et seulement si $\det(A) \neq 0$. Dans ce cas, sa réciproque est encore affine :

$$w^{-1}\begin{pmatrix} x \\ y \end{pmatrix} = A^{-1}\begin{pmatrix} x \\ y \end{pmatrix} - A^{-1}\mathbf{t}.$$

Démonstration. Si $\det(A) \neq 0$, la matrice A est inversible, et on peut vérifier que l'application qui envoie $\begin{pmatrix} x \\ y \end{pmatrix}$ sur $A^{-1}\left(\begin{pmatrix} x \\ y \end{pmatrix} - \mathbf{t}\right)$ est une réciproque de w : cette réciproque est affine, avec une partie linéaire A^{-1} et de partie translation $-A^{-1}\mathbf{t}$. Réciproquement, si $\det(A) = 0$, la matrice A possède un vecteur non nul $\mathbf{u}$ dans son noyau ; alors $w(P + \mathbf{u}) = w(P)$ pour tout point P, donc w n'est pas injective. □

Ainsi, les applications affines inversibles constituent un sous-groupe de $\mathcal{T}$. Nous le noterons $\mathcal{A}$ et l'appellerons le **groupe affine** du plan.

Nous pouvons maintenant préciser une convention de vocabulaire qui sera utilisée dans tout le chapitre. Une **application affine** est n'importe quelle fonction de la forme (1.1), sans condition sur A : la projection orthogonale de la section 1.6 en est une. Une **transformation affine** est une application affine *bijective*, c'est-à-dire un élément de $\mathcal{A}$. Le mot *transformation* garde ainsi le sens que lui donne la définition 1.1 : l'inversibilité est déjà comprise dans l'expression « transformation affine ». Une application affine non inversible n'est donc pas une transformation au sens retenu ici.

1.5.1 Ce qu'une transformation affine conserve

Proposition 1.3 **Propriétés des transformations affines**

Soit w une transformation affine du plan. Alors :

1. l'image d'une droite est une droite ;
2. l'image d'une droite passant par l'origine est une droite passant par (e, f) ;
3. l'image de deux droites parallèles est formée de deux droites parallèles ;
4. l'image d'un segment est un segment, et l'image du milieu d'un segment est le milieu du segment image ;
5. plus généralement, w préserve les rapports de longueurs *sur une même droite* ;
6. l'image d'un carré est un parallélogramme, et l'image d'un parallélogramme est un parallélogramme.

Démonstration. Utilisons la représentation paramétrique d'une droite, vue en géométrie vectorielle : une **droite** passant par P et de **vecteur directeur** $\mathbf{u} \neq \mathbf{0}$ (un vecteur non nul parallèle à la droite) est l'ensemble des points $P + s\mathbf{u}$, $s \in \mathbb{R}$: on obtient tous les points de la droite en partant de P et en avançant d'un multiple quelconque de $\mathbf{u}$. En appliquant

la transformation, nous obtenons

$$w(P + s\mathbf{u}) = A(P + s\mathbf{u}) + \mathbf{t} = \underbrace{AP + \mathbf{t}}_{= w(P)} + s\,(A\mathbf{u}),$$

c'est-à-dire la droite passant par $w(P)$ et de vecteur directeur $A\mathbf{u}$, lequel est non nul puisque A est inversible. Cela démontre (1), et (2) en prenant $P = \mathbf{0}$.

Pour établir (3), observons que deux droites parallèles ont le même vecteur directeur $\mathbf{u}$, donc leurs images ont le même vecteur directeur $A\mathbf{u}$; elles sont parallèles.

Pour (4) et (5), rappelons que le segment $[PQ]$ est l'ensemble des $P + s(Q - P)$ pour $s \in [0, 1]$, et le calcul précédent montre que son image est l'ensemble des $w(P) + s\,A(Q - P)$ pour les mêmes valeurs de s, soit le segment $[w(P)\,w(Q)]$. Puisque la valeur du paramètre s est conservée, le point qui coupait le segment dans le rapport s est envoyé sur le point qui coupe le segment image dans le même rapport s. En particulier le milieu ($s = 1/2$) va sur le milieu.

Enfin, pour (6), le carré de sommets P, $P+\mathbf{u}$, $P+\mathbf{u}+\mathbf{v}$, $P+\mathbf{v}$ a pour image le quadrilatère de sommets $w(P)$, $w(P) + A\mathbf{u}$, $w(P) + A\mathbf{u} + A\mathbf{v}$, $w(P) + A\mathbf{v}$, dont les côtés opposés sont parallèles deux à deux. □

Remarque 1.6 **Les propriétés qui ne sont pas conservées**

Il faut aussi préciser les limites de ce résultat : une transformation affine ne préserve en général ni les longueurs, ni les angles, ni la perpendicularité, ni les rapports de longueurs entre deux directions différentes. Un carré peut devenir un parallélogramme quelconque ; un cercle devient une ellipse. En revanche, elle multiplie toutes les aires par le même facteur, comme nous allons le voir.

1.5.2 Visualiser une transformation affine : l'image du carré unité

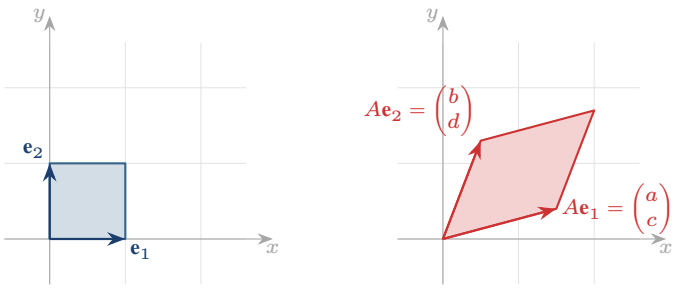
Une façon de comprendre l'effet d'une transformation affine consiste à examiner l'image du **carré unité** $[0, 1] \times [0, 1]$. Cette observation est aussi une activité utile pour aborder les transformations avec des élèves.

Pour comprendre pourquoi, notons $\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ et $\mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. Alors

$$A\mathbf{e}_1 = \begin{pmatrix} a \\ c \end{pmatrix}, \qquad A\mathbf{e}_2 = \begin{pmatrix} b \\ d \end{pmatrix}.$$

On constate donc que *les colonnes de A sont les images des deux vecteurs de la base*. Le carré unité, dont les côtés sont $\mathbf{e}_1$ et $\mathbf{e}_2$, devient donc le parallélogramme construit sur les deux colonnes de A, puis translaté de $\mathbf{t}$.

La proposition suivante fait intervenir l'**orientation** du plan. Avant de l'énoncer, précisons le sens de cette notion.

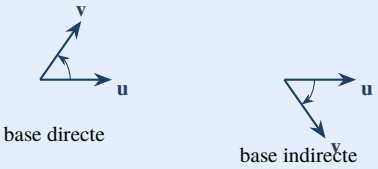


le carré unité — son image par la matrice A

FIGURE 1.4 : Les colonnes de la matrice sont les images des vecteurs de la base. Le carré unité devient le parallélogramme qu'elles engendrent. Cette figure permet ainsi de lire les coefficients de A.

Définition 1.4 **Orientation d'une base, du plan**

Une base $(\mathbf{u}, \mathbf{v})$ du plan est dite **directe** si l'on passe du premier vecteur au second, en tournant par le plus petit angle, dans le sens antihoraire (le sens trigonométrique) ; elle est dite **indirecte** dans le cas contraire.



base directe — base indirecte

On dit qu'une transformation affine w **préserve l'orientation** si elle envoie toute base directe sur une base directe (et donc, nécessairement, toute base indirecte sur une base indirecte) ; on dit qu'elle la **renverse** si elle échange les deux types de bases. Puisque A envoie la base standard directe $(\mathbf{e}_1, \mathbf{e}_2)$ sur $(A\mathbf{e}_1, A\mathbf{e}_2)$, et que cette dernière est directe exactement lorsque $\det(A) > 0$, le signe du déterminant décide de tout, comme l'affirme la proposition suivante.

Proposition 1.4 **Aire et déterminant**

Soit w une application affine de partie linéaire A, et soit $D \subset \mathbb{R}^2$ une région d'aire S. Alors

$$\text{Aire}(w(D)) = |\det(A)| \cdot S.$$

Lorsque $\det(A) \neq 0$, c'est-à-dire lorsque w est une transformation affine, elle préserve de plus l'orientation du plan si $\det(A) > 0$ et la renverse si $\det(A) < 0$.

Nous admettrons ce résultat, qui peut être démontré à l'aide du théorème du changement de variables en calcul intégral. On peut cependant en expliquer l'idée géométrique. La propriété est vérifiée pour le carré unité, par la définition du déterminant, puis pour les autres carrés d'un quadrillage, par translation. Elle s'étend aux réunions de ces carrés. En utilisant un quadrillage de plus en plus fin, on peut ensuite approcher l'aire d'une région plus générale.

Pour la classe 1.2 **Le F est-il à l'endroit ?**

Les déterminants ne sont pas étudiés au secondaire, mais la symétrie axiale l'est. On peut donc s'appuyer sur cette notion familière pour interpréter le signe du

déterminant : si $\det(A) > 0$, la figure image se lit dans le même sens que la figure de départ ; si $\det(A) < 0$, elle est retournée, comme dans un miroir. Pour rendre ce retournement visible, nous utilisons un motif en forme de « F », qui n'a pas de symétrie axiale. Un triangle isocèle, au contraire, possède une symétrie intrinsèque qui pourrait masquer le changement d'orientation.

Outil 1.1 **Une matrice à curseurs dans GeoGebra**

Dans GeoGebra, on crée quatre curseurs a, b, c et d, la matrice `M = {{a, b}, {c, d}}` et un polygone en forme de « F », puis on trace son image avec la commande `ApplyMatrix(M, F)`. En affichant aussi `Determinant(M)`, les élèves peuvent observer en direct ce que les résultats de cette section prédisent : le rapport des aires suit $|\det(M)|$, le « F » se retourne au moment où le déterminant change de signe, et la figure s'aplatit sur une droite lorsqu'il s'annule (section 1.6.8). Pour éviter que l'activité se réduise à faire bouger des curseurs, on demandera d'abord de *prédire* l'effet d'un réglage (remarque 1.11).

Exemple 1.5 **Image d'un rectangle**

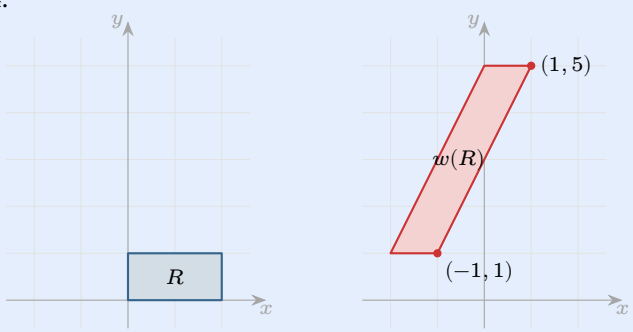
Déterminons l'image du rectangle plein $R = [0, 2] \times [0, 1]$ par

$$w\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 2 & 0 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

D'après la proposition 1.3, l'image d'un parallélogramme est un parallélogramme. Nous pouvons donc déterminer l'image du rectangle en calculant celles de ses quatre sommets. On obtient

$$\begin{aligned} w(0,0) &= (-1,1), & w(2,0) &= (1,5), \\ w(2,1) &= (0,5), & w(0,1) &= (-2,1). \end{aligned}$$

L'image est donc le parallélogramme plein de sommets $(-1, 1)$, $(1, 5)$, $(0, 5)$ et $(-2, 1)$. Vérifions aussi le résultat en comparant les aires : $\det(A) = 0 - (-2) = 2$, et l'aire du rectangle valait 2 ; l'aire de l'image doit donc valoir 4. C'est bien le cas : le parallélogramme a une base de longueur 1 (le segment de $(-1, 1)$ à $(-2, 1)$) et une hauteur 4.



Exemple 1.6 **Image d'un triangle**

Avec la même transformation w, cherchons l'image du triangle plein de sommets $A = (1,1)$, $B = (0,0)$ et $C = (-1,1)$. On calcule

$$w(1,1) = \begin{pmatrix} 1-1-1 \\ 2+0+1 \end{pmatrix} = \begin{pmatrix} -1 \\ 3 \end{pmatrix}, \quad w(0,0) = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \quad w(-1,1) = \begin{pmatrix} -3 \\ -1 \end{pmatrix}.$$

L'image est le triangle plein de sommets $(-1,3)$, $(-1,1)$ et $(-3,-1)$. Observons que le triangle de départ est isocèle, alors que son image ne l'est pas. Cet exemple rappelle qu'une transformation affine ne conserve pas nécessairement les longueurs.

1.6 Un catalogue d'applications affines

Nous allons présenter quelques applications affines élémentaires. À l'exception de la projection, étudiée en dernier, elles sont toutes des transformations du plan. Leur intérêt vient notamment du fait que toute application linéaire inversible du plan peut être obtenue en les composant, comme nous le verrons au théorème 1.7.

Pour la suite, nous désignerons l'origine par $\mathbf{0}$ et représenterons un point du plan par le vecteur colonne $\begin{pmatrix} x \\ y \end{pmatrix}$.

Nous adopterons les conventions de notation suivantes dans tout le chapitre : les *rotations* portent un r minuscule et les *réflexions* un R majuscule. Ainsi r_θ est la rotation d'angle θ autour de l'origine et $r_{O,\alpha}$ celle d'angle α autour du point O, tandis que R_θ est la réflexion dont l'axe passe par l'origine en faisant un angle θ avec l'axe des x, et R_d la réflexion d'axe d. Les translations s'écrivent $t_{\mathbf{t}}$ et les homothéties h_c ou $h_{M,c}$.

1.6.1 Translations

Définition 1.5 **Translation**

La **translation** de vecteur $\mathbf{t} = \begin{pmatrix} e \\ f \end{pmatrix}$ est l'application

$$t_{\mathbf{t}} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix}.$$

Sa partie linéaire est l'identité : $\det(I) = 1$, donc les aires et l'orientation sont préservées. La translation fournit ainsi un premier exemple d'isométrie. Notons aussi que composer deux translations donne une translation, et que l'ordre n'y change rien : $t_{\mathbf{u}} \circ t_{\mathbf{v}} = t_{\mathbf{u}+\mathbf{v}} = t_{\mathbf{v}} \circ t_{\mathbf{u}}$. Dans ce cas particulier, l'ordre de composition n'a donc pas d'importance. Cette simplicité peut aider à comprendre la place accordée aux translations au début du secondaire.

1.6.2 Homothéties

Définition 1.6 **Homothétie de centre l'origine**

L'**homothétie de centre 0 et de rapport** $c \in \mathbb{R}^\star$ est l'application linéaire

$$h_c \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} c & 0 \\ 0 & c \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} cx \\ cy \end{pmatrix}.$$

Lorsque $0 < c < 1$, on parle de **contraction uniforme** ; lorsque $c > 1$, de **dilatation uniforme**. Lorsque $c < 0$, l'homothétie retourne la figure de l'autre côté du centre ; en particulier h_{-1} est la **symétrie centrale** de centre $\mathbf{0}$, qui n'est autre que la rotation d'un demi-tour.

On a $\det(cI) = c^2 > 0$: une homothétie multiplie les aires par c^2 et ne renverse jamais l'orientation, même si $c < 0$. Le cas d'un rapport négatif mérite d'être discuté en classe, car il peut sembler contraire à l'intuition.

Pour décrire une homothétie dont le centre n'est pas l'origine, nous utiliserons la **conjugaison**. Le procédé consiste à translater le centre à l'origine, à appliquer l'homothétie déjà définie, puis à effectuer la translation inverse.

Proposition 1.5 **Homothétie de centre quelconque**

L'homothétie de centre $M = (h, k)$ et de rapport c est

$$h_{M,c} = t_{\mathbf{m}} \circ h_c \circ t_{-\mathbf{m}}, \qquad \mathbf{m} = \begin{pmatrix} h \\ k \end{pmatrix},$$

et sa représentation analytique est

$$h_{M,c} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} c & 0 \\ 0 & c \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} h(1-c) \\ k(1-c) \end{pmatrix} = \begin{pmatrix} cx + h(1-c) \\ cy + k(1-c) \end{pmatrix}.$$

Démonstration. Notons P le point de départ et appliquons successivement les trois transformations : $P \mapsto P - \mathbf{m} \mapsto c(P - \mathbf{m}) \mapsto c(P - \mathbf{m}) + \mathbf{m} = cP + (1 - c)\mathbf{m}$.

Vérifions que cette composée satisfait la définition 1.6 avec centre M plutôt qu'avec centre $\mathbf{0}$. Or, en soustrayant $\mathbf{m}$ de part et d'autre de $h_{M,c}(P) = cP + (1 - c)\mathbf{m}$, on obtient

$$h_{M,c}(P) - M = c(P - M),$$

c'est-à-dire que le vecteur allant de M à l'image est c fois le vecteur allant de M à P. Cela signifie que l'image P' de P est sur la droite MP, du même côté de M que P si $c > 0$ et du côté opposé si $c < 0$, à une distance $|c|$ fois plus grande. C'est la définition synthétique de l'homothétie (exemple 1.1 pour le cas $c = -1$), transportée sans changement au centre M. De plus, le centre M est fixe, puisque $h_{M,c}(M) - M = c(M - M) = \mathbf{0}$. □

Pour la classe 1.3 **Interpréter la formule de l'homothétie**

La formule $(x, y) \mapsto (cx + h(1-c),\ cy + k(1-c))$ se retrouve dans les manuels de deuxième secondaire. À ce niveau, elle est généralement présentée sans le calcul qui permet de l'obtenir. La composition $t_{\mathbf{m}} \circ h_c \circ t_{-\mathbf{m}}$ en explique pourtant la forme : on déplace le centre à l'origine, on effectue l'homothétie, puis on revient par la translation inverse. Ce procédé permet de reconstruire les formules des transformations à centre étudiées au secondaire. Il est donc utile de pouvoir le

reprendre au tableau, même si on n'en présente pas tout le formalisme aux élèves.

1.6.3 Rotations

Définition 1.7 **Rotation de centre l'origine**

La **rotation** d'angle θ autour de l'origine, dans le sens antihoraire, est l'application linéaire de matrice

$$r_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

Remarque 1.7 **Un ou deux indices ?**

Précisons la lecture des indices, conformément à la convention de la section 1.6 : un indice numérique seul, comme dans r_θ, signifie que la rotation est centrée à l'*origine*. Dès que le centre n'est pas l'origine, on le précise en le plaçant en premier indice, comme dans $r_{M,\theta}$, introduit plus bas : l'indice porte alors sur *deux* objets, le centre et l'angle, et non seulement l'angle. La même lecture vaudra pour les homothéties (h_c contre $h_{M,c}$) et pour les réflexions (R_θ contre R_d).

On peut retrouver cette matrice à partir de la méthode présentée au paragraphe 1.5.2 : la première colonne est $(\cos\theta, \sin\theta)$, qui est bien l'image de $\mathbf{e}_1$ par une rotation d'angle θ ; la seconde est $(-\sin\theta, \cos\theta)$, soit $\mathbf{e}_2$ tourné du même angle. On a $\det(r_\theta) = \cos^2\theta + \sin^2\theta = 1$: les rotations préservent les aires et l'orientation.

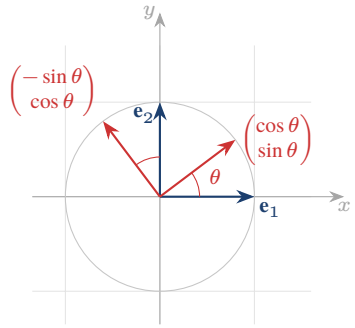


FIGURE 1.5 : Les deux colonnes de r_θ sont les images de $\mathbf{e}_1$ et $\mathbf{e}_2$. La lecture des coordonnées de ces deux points du cercle trigonométrique permet de reconstruire la matrice.

En reprenant le procédé de conjugaison de la proposition 1.5, on obtient la rotation de centre M quelconque :

$$r_{M,\theta}\begin{pmatrix} x \\ y \end{pmatrix} = r_\theta\left(\begin{pmatrix} x \\ y \end{pmatrix} - \mathbf{m}\right) + \mathbf{m} = r_\theta\begin{pmatrix} x \\ y \end{pmatrix} + (I - r_\theta)\,\mathbf{m}.$$

Proposition 1.6 **Composer deux rotations de même centre**

Soit M un point quelconque du plan. Alors $r_{M,\alpha}\, r_{M,\beta} = r_{M,\alpha+\beta} = r_{M,\beta}\, r_{M,\alpha}$, et $r_{M,\theta}^{-1} = r_{M,-\theta}$.

Démonstration. Traitons d'abord le cas $M = \mathbf{0}$, c'est-à-dire les rotations r_α et r_β de la définition 1.7. Le calcul du produit matriciel fait intervenir les formules d'addition des

angles :

$$\begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} \begin{pmatrix} \cos\beta & -\sin\beta \\ \sin\beta & \cos\beta \end{pmatrix} = \begin{pmatrix} p & -q \\ q & p \end{pmatrix},$$

où

$$\begin{aligned} p &= \cos\alpha\cos\beta - \sin\alpha\sin\beta = \cos(\alpha+\beta), \\ q &= \sin\alpha\cos\beta + \cos\alpha\sin\beta = \sin(\alpha+\beta). \end{aligned}$$

On reconnaît $r_{\alpha+\beta}$, et le produit dans l'autre ordre donne le même résultat puisque $\alpha+\beta = \beta+\alpha$: c'est $r_\alpha r_\beta = r_{\alpha+\beta} = r_\beta r_\alpha$.

Pour un centre M quelconque, de vecteur position $\mathbf{m}$, la conjugaison introduite ci-dessus donne, pour tout point P,

$$\begin{aligned} (r_{M,\alpha}\, r_{M,\beta})(P) &= r_\alpha(r_\beta(P-\mathbf{m})+\mathbf{m}-\mathbf{m})+\mathbf{m} \\ &= (r_\alpha r_\beta)(P-\mathbf{m})+\mathbf{m} \\ &= r_{\alpha+\beta}(P-\mathbf{m})+\mathbf{m} \\ &= r_{M,\alpha+\beta}(P), \end{aligned}$$

où l'on a utilisé le cas $M = \mathbf{0}$ à la troisième égalité. En inversant l'ordre des rotations, le calcul précédent donne le même résultat, et il suffit de poser $\beta = -\alpha$ pour obtenir $r_{M,\theta}^{-1} = r_{M,-\theta}$. □

Pour la classe 1.4 **Les identités trigonométriques et les rotations**

La démonstration ci-dessus est une preuve des identités $\cos(\alpha+\beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta$ et $\sin(\alpha+\beta) = \sin\alpha\cos\beta + \cos\alpha\sin\beta$, à condition d'admettre que composer deux rotations donne une rotation, ce qui est géométriquement évident. Cette démonstration permet de comprendre la forme de ces identités en les reliant au produit de deux matrices. Elle peut ainsi enrichir leur présentation en cinquième secondaire.

1.6.4 Réflexions

Définition 1.8 **Réflexions par rapport aux axes**

Les réflexions par rapport à l'axe des x et à l'axe des y sont respectivement

$$H = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \qquad V = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Plus généralement, la réflexion par rapport à la droite passant par l'origine et faisant un angle θ avec l'axe des x a pour matrice

$$R_\theta = \begin{pmatrix} \cos 2\theta & \sin 2\theta \\ \sin 2\theta & -\cos 2\theta \end{pmatrix}. \tag{1.2}$$

On vérifie que $\det(R_\theta) = -\cos^2 2\theta - \sin^2 2\theta = -1$: une réflexion renverse l'orientation, comme le suggère l'image du miroir. Examinons deux cas particuliers : $\theta = 0$ donne H ; $\theta = \pi/4$ donne la réflexion par rapport à la droite $y = x$, de matrice $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, qui échange les coordonnées.

Remarque 1.8 **L'angle double**

La présence de 2θ dans (1.2) s'explique géométriquement : faire tourner le *miroir* de θ fait tourner l'*image* de 2θ. C'est le même phénomène qui fait que la composée de deux réflexions d'axes sécants est une rotation d'angle *double* de l'angle des axes, résultat que nous démontrons au théorème 1.13 et que l'on peut faire observer aux élèves du secondaire.

1.6.5 Cisaillements

Définition 1.9 **Cisaillement**

Un **cisaillement horizontal** de rapport m est l'application linéaire de matrice $\begin{pmatrix} 1 & m \\ 0 & 1 \end{pmatrix}$; un **cisaillement vertical** de rapport m, celle de matrice $\begin{pmatrix} 1 & 0 \\ m & 1 \end{pmatrix}$.

Un cisaillement horizontal laisse fixe l'axe des x et fait glisser chaque droite horizontale d'une quantité proportionnelle à sa hauteur : on peut visualiser cet effet en décalant progressivement les cartes d'un paquet initialement bien empilé. Son déterminant vaut 1 : *un cisaillement conserve les aires*, bien que les angles et les longueurs puissent être modifiés.

Pour la classe 1.5 **Le cisaillement et l'aire du parallélogramme**

Le cisaillement permet de justifier la formule « base $\times$ hauteur » pour l'aire d'un parallélogramme. En effet, un cisaillement de rapport approprié transforme celui-ci en un rectangle de même base et de même hauteur. Puisque l'aire est conservée, les deux figures ont la même aire. Au secondaire, on retrouve cette égalité d'aires en découpant un triangle et en le déplaçant de l'autre côté , ce qui donne une autre interprétation du résultat.

1.6.6 Contractions et dilatations non uniformes

Définition 1.10 **Contraction non uniforme**

C'est une application linéaire de la forme $\begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}$ avec $0 < a < 1$ et $0 < b < 1$.

Si $a \neq b$, les facteurs de contraction diffèrent suivant les deux directions : un cercle devient une ellipse.

1.6.7 Un catalogue en images

1.6.8 Le cas non inversible

Il est parfois utile de considérer des applications affines qui ne sont pas des transformations, c'est-à-dire pour lesquelles $\det(A) = 0$. L'image du plan est alors une droite (si $A \neq 0$) ou un point (si $A = 0$).

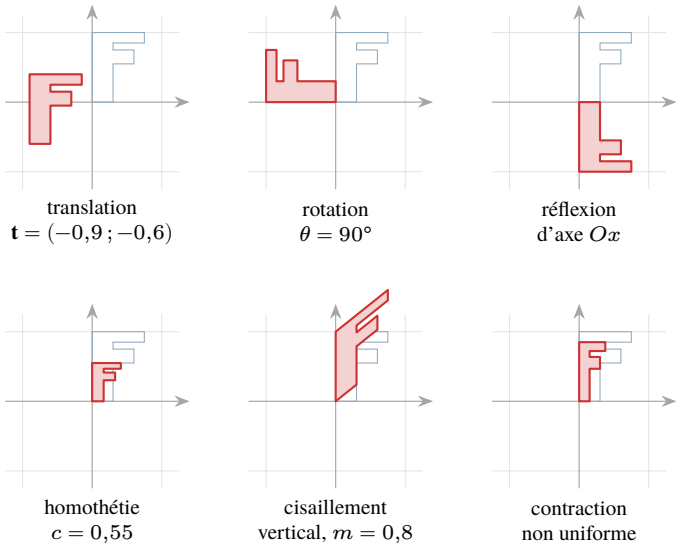


FIGURE 1.6 : Quelques applications affines, représentées à l'aide d'un même motif. La figure de départ est en trait pâle et son image en couleur. Parmi les transformations illustrées, seule la réflexion renverse l'orientation, ce que permet de constater la forme du « F ».

TABLE 1.1 : Les applications affines élémentaires et leurs propriétés. La projection est la seule de cette liste qui ne soit pas une transformation du plan, puisque son déterminant est nul.

Transformation	Matrice A	$\det(A)$	Isométrie ?
Translation de $\mathbf{t}$	I	1	oui
Rotation d'angle θ	$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$	1	oui
Réflexion d'axe θ	$\begin{pmatrix} \cos 2\theta & \sin 2\theta \\ \sin 2\theta & -\cos 2\theta \end{pmatrix}$	-1	oui
Homothétie de rapport c	$\begin{pmatrix} c & 0 \\ 0 & c \end{pmatrix}$	c^2	si $\lvert c\rvert = 1$
Cisaillement de rapport m	$\begin{pmatrix} 1 & m \\ 0 & 1 \end{pmatrix}$	1	non (sauf $m = 0$)
Contraction non uniforme	$\begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}$	ab	non (sauf $\lvert a\rvert = \lvert b\rvert = 1$)
Projection sur $y = x$	$\frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$	0	non (non inversible)

Exemple 1.7 **La projection orthogonale sur $y = x$**

Après le repliage de l'exemple 1.2, considérons une autre application non inversible, qui est cette fois affine : la **projection orthogonale** sur une droite. La projection orthogonale du point (x, y) sur la droite $y = x$ est le point de cette droite le plus proche de (x, y) ; en notant $\mathbf{u} = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix}$ le vecteur unitaire directeur, c'est

$$p\begin{pmatrix}x\\y\end{pmatrix} = \left(\begin{pmatrix}x\\y\end{pmatrix}\cdot\mathbf{u}\right)\mathbf{u} = \frac{x+y}{2}\begin{pmatrix}1\\1\end{pmatrix} = \frac{1}{2}\begin{pmatrix}1&1\\1&1\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}.$$

Le déterminant vaut $\frac{1}{4}(1-1) = 0$: les images ont une aire nulle, ce qui correspond au fait qu'elles sont contenues dans une droite. Observons aussi que $p \circ p = p$: projeter deux fois revient à projeter une fois.

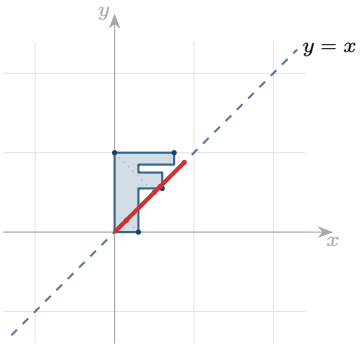


FIGURE 1.7 : Une application affine non inversible : l'image du motif est le segment de la droite $y = x$ allant de $(0, 0)$ à $(\frac{7}{8}, \frac{7}{8})$, image du coin $(\frac{3}{4}, 1)$. Ce n'est pas une transformation du plan au sens de la définition 1.1.

1.6.9 Le théorème de décomposition

Théorème 1.7 **Décomposition des applications linéaires du plan**

Une application linéaire de $\mathbb{R}^2$ dans $\mathbb{R}^2$ est inversible si et seulement si elle est une composition d'homothéties, de rotations, de réflexions, de cisaillements et de contractions ou dilatations non uniformes.

On peut interpréter ce résultat à partir d'une propriété de l'algèbre linéaire : toute matrice inversible est un produit de matrices élémentaires, celles qui correspondent aux opérations sur les lignes. Les cisaillements sont exactement les matrices d'addition d'un multiple d'une ligne à une autre ; les contractions non uniformes, celles de multiplication d'une ligne par un scalaire ; la réflexion d'axe $y = x$, celle d'échange de deux lignes. Ainsi, la méthode de Gauss–Jordan, généralement présentée au cégep comme une technique de calcul, possède aussi une interprétation géométrique : les opérations élémentaires correspondent à une *décomposition* de la transformation du plan.

1.7 Les coordonnées homogènes et leurs applications en infographie

La composition de deux applications affines fait intervenir à la fois leurs parties linéaires et leurs translations. Si w_1 et w_2 ont pour parties linéaires A_1 et A_2 et pour translations $\mathbf{t}_1$ et $\mathbf{t}_2$, alors

$$(w_2 \circ w_1)\begin{pmatrix} x \\ y \end{pmatrix} = A_2\left(A_1\begin{pmatrix} x \\ y \end{pmatrix} + \mathbf{t}_1\right) + \mathbf{t}_2 = (A_2A_1)\begin{pmatrix} x \\ y \end{pmatrix} + (A_2\mathbf{t}_1 + \mathbf{t}_2),$$

Cette écriture demande donc deux calculs distincts. Lorsqu'on doit composer un grand nombre de transformations, comme en infographie, il est utile de disposer d'une représentation qui réunisse ces opérations.

Pour y parvenir, on plonge le plan dans $\mathbb{R}^3$, à la hauteur $z = 1$: on représente le point (x, y) par le triplet $(x, y, 1)$. On écrit alors

$$\begin{pmatrix} a & b & e \\ c & d & f \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} ax + by + e \\ cx + dy + f \\ 1 \end{pmatrix}. \tag{1.3}$$

Le terme de translation est maintenant compris dans le produit matriciel. L'application affine du plan est ainsi représentée par une application linéaire dans un espace de dimension supérieure. On appelle $(x, y, 1)$ les **coordonnées homogènes** du point (x, y).

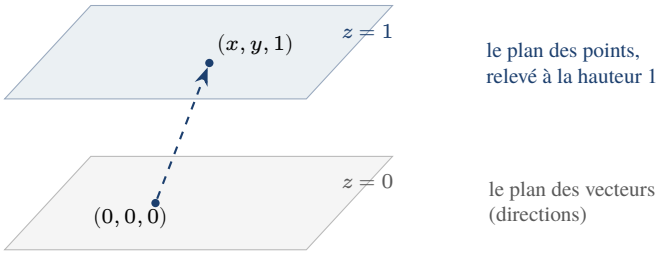


FIGURE 1.8 : Les coordonnées homogènes. Les points du plan sont représentés à la hauteur $z = 1$, les vecteurs (les directions) à la hauteur $z = 0$. La distinction point/vecteur de la remarque 1.4 apparaît dans la troisième coordonnée.

Cette représentation possède plusieurs propriétés utiles :

1. **Composition.** La matrice de $w_2 \circ w_1$ est le produit des deux matrices 3×3, dans cet ordre. Le produit tient compte à la fois de la partie linéaire et de la translation.
2. **Inversion.** La matrice de w^{-1} est l'inverse de la matrice 3×3.
3. **Points et vecteurs.** Un point est $(x, y, 1)$, un vecteur est $(x, y, 0)$. La différence de deux points donne bien un vecteur, et la translation d'un vecteur ne le change pas : c'est exactement ce que dit la géométrie affine.
4. **Extension aux transformations projectives.** En permettant à la troisième ligne d'être autre chose que $(0, 0, 1)$, on obtient les transformations *projectives*, celles qui font converger les parallèles vers un point de fuite. C'est ainsi que fonctionne toute perspective en trois dimensions [Fol+96].

Exemple 1.8 **Composer une rotation et une translation**

Considérons le motif en forme de maison de sommets $(0,0)$, $(2,0)$, $(2\,;\,1{,}4)$, $(1\,;\,2{,}2)$, $(0\,;\,1{,}4)$, et appliquons successivement deux transformations : r, la rotation de $\pi/2$ autour de l'origine, puis t, la translation de vecteur $(3,1)$. En coordonnées homogènes,

$$M_r = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad M_t = \begin{pmatrix} 1 & 0 & 3 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

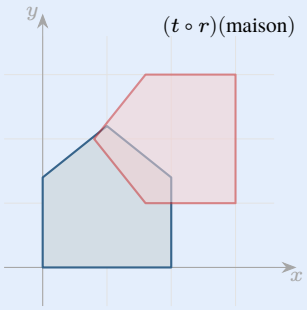
Puisque la rotation précède la translation, la transformation composée $t \circ r$ a pour matrice le produit $M_t M_r$, calculé dans cet ordre :

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

Appliquons-la, par exemple, au coin $(2,0)$ de la maison, écrit en coordonnées homogènes $(2,0,1)$:

$$M_t M_r \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0(2) - 1(0) + 3(1) \\ 1(2) + 0(0) + 1(1) \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \\ 1 \end{pmatrix},$$

c'est-à-dire $(2,0) \mapsto (3,3)$. C'est exactement ce que donnerait le calcul affine direct $\begin{pmatrix} x \\ y \end{pmatrix} \mapsto r_{\pi/2} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 3 \\ 1 \end{pmatrix}$, mais obtenu ici par une seule multiplication de matrices 3×3.



maison

Composons maintenant dans l'autre ordre, $r \circ t$ (la translation précède alors la rotation) :

$$M_r M_t = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 3 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & -1 & -1 \\ 1 & 0 & 3 \\ 0 & 0 & 1 \end{pmatrix},$$

On obtient une matrice différente de $M_t M_r$. Le recours aux coordonnées homogènes ne change donc pas l'importance de l'ordre de composition, déjà soulignée à la remarque 1.1. Nous retrouverons ces deux mêmes matrices, calculées cette fois par la machine, à l'annexe A.

Remarque 1.9 **Quelques applications**

Les produits de matrices de la forme (1.3) interviennent dans de nombreuses applications : déplacer un personnage dans un jeu vidéo, déterminer la position de la pince d'un bras de robot articulé [Cra05], redresser une photographie ou faire pivoter une carte. Le nombre de calculs peut atteindre des millions par seconde. Ces exemples permettent de montrer aux élèves comment des transformations rencontrées au secondaire sont utilisées dans les technologies qu'ils connaissent.

1.8 Isométries et similitudes

Parmi les transformations affines, certaines préservent la *forme* des figures. Nous allons les étudier et préciser leur place dans la hiérarchie de la figure 1.3. Ce sont notamment les transformations auxquelles on s'intéresse au secondaire.

Définition 1.11 **Similitude, isométrie, contraction**

Une **similitude** du plan est une application affine $w : \mathbb{R}^2 \to \mathbb{R}^2$ pour laquelle il existe une constante $k > 0$ telle que

$$\|w(P) - w(Q)\| = k\,\|P - Q\| \qquad \text{pour tous } P, Q \in \mathbb{R}^2\,. \tag{1.4}$$

Le nombre k est le **rapport de similitude**. Lorsque $k = 1$, on dit que w est une **isométrie**. Lorsque $0 < k < 1$, on dit que w est une **contraction** de facteur k.

La condition (1.4) signifie que toutes les distances sont multipliées par le même facteur k. On retrouve la propriété utilisée au secondaire pour définir des figures semblables. Cependant, elle porte ici sur une *transformation* du plan, plutôt que sur une paire de figures donnée.

Proposition 1.8 **Il suffit de regarder la partie linéaire**

Une application affine de partie linéaire A est une similitude de rapport k si et seulement si l'application linéaire $\begin{pmatrix} x \\ y \end{pmatrix} \mapsto A \begin{pmatrix} x \\ y \end{pmatrix}$ en est une, de même rapport.

Démonstration. Observons d'abord que le terme de translation disparaît dans la différence :

$$w(P) - w(Q) = (AP + \mathbf{t}) - (AQ + \mathbf{t}) = A(P - Q) = w_1(P) - w_1(Q).$$

Les deux membres de (1.4) sont donc identiques pour w et pour w_1. □

Proposition 1.9 **Une composition de similitudes est une similitude**

Si w_1 et w_2 sont des similitudes de rapports k_1 et k_2, alors $w_2 \circ w_1$ est une similitude de rapport $k_1 k_2$.

Démonstration. Pour tous $P, Q \in \mathbb{R}^2$,

$$\|w_2(w_1(P)) - w_2(w_1(Q))\| = k_2\,\|w_1(P) - w_1(Q)\| = k_2 k_1 \|P - Q\|.$$

Une récurrence permet ensuite d'étendre cette égalité à un nombre quelconque de facteurs. □

En particulier, les similitudes forment un sous-groupe $\mathcal{S}$ de $\mathcal{A}$, et les isométries un sous-groupe $\mathcal{I}$ de $\mathcal{S}$: ce qui justifie les inclusions représentées à la figure 1.3.

1.8.1 Les propriétés des similitudes

Nous savons que les homothéties, les rotations, les réflexions et les translations sont des similitudes (exercice 1.6), de même que toutes leurs composées. Nous admettrons la réciproque, dont la démonstration dépasse le cadre du cours : la partie linéaire d'une similitude de rapport k est toujours de la forme $k\,r_\theta$ (similitude directe) ou $k\,R_\theta$ (similitude indirecte) pour un certain angle θ. Les deux propositions suivantes présentent des conséquences de cette caractérisation.

Proposition 1.10 **Déterminant d'une similitude**

Si w est une similitude de rapport k et de partie linéaire A, alors $\det(A) = k^2$ ou $\det(A) = -k^2$. Par conséquent $|\det(A)| = k^2$, et w est une contraction si et seulement si $|\det(A)| < 1$.

Proposition 1.11 **Aires et rapport de similitude**

Si w est une similitude de rapport k et si D est une région du plan, alors $\text{Aire}(w(D)) = k^2\ \text{Aire}(D)$.

Pour la classe 1.6 **Le facteur k^2 pour les aires**

Au secondaire, la proposition 1.11 s'exprime souvent ainsi : « si on double les dimensions, l'aire est multipliée par quatre ». La mémorisation de cette règle ne suffit pas toujours à bien l'appliquer. Pour en comprendre la raison, on peut proposer deux explications : l'une par le déterminant (proposition 1.4), l'autre par le fait qu'une aire est un produit de deux longueurs, chacune multipliée par k. La deuxième est parfaitement accessible à un élève de troisième secondaire ; la première explique pourquoi le résultat vaut aussi pour des figures qui ne sont ni des rectangles ni des triangles. Le même raisonnement en dimension trois donne k^3 pour les volumes.

Exemple 1.9 **Reconnaître une similitude**

La matrice $A = \begin{pmatrix} 0{,}8 & 0 \\ a & 0{,}6 \end{pmatrix}$ ne représente une similitude pour aucune valeur de a. En effet, si k était le rapport, l'exercice 1.5 appliqué à $\mathbf{e}_1$ et $\mathbf{e}_2$ donnerait

$$k = \|A\mathbf{e}_1\| = \sqrt{0{,}8^2 + a^2} \geq 0{,}8 \qquad \text{et} \qquad k = \|A\mathbf{e}_2\| = 0{,}6,$$

ce qui est contradictoire.

En revanche, $A = \begin{pmatrix} 0{,}5 & \alpha \\ \beta & 0{,}5 \end{pmatrix}$ est une similitude pour $\alpha = \beta = 0$ (homothétie de rapport $\frac{1}{2}$) et aussi pour $\alpha = \frac{\sqrt{3}}{2}$, $\beta = -\frac{\sqrt{3}}{2}$: dans ce dernier cas, on reconnaît la matrice $r_{-\pi/3}$ de la rotation d'angle $-\pi/3$: c'est même une isométrie, puisque $k = 1$.

1.8.2 Classification des isométries du plan

Théorème 1.12 **Les quatre isométries du plan**

Toute isométrie du plan est de l'un des quatre types suivants :

1. une **translation** (y compris l'identité) ;
2. une **rotation** autour d'un point ;
3. une **réflexion** par rapport à une droite ;
4. une **réflexion glissée** (ou **symétrie glissée** : les deux noms coexistent dans la littérature — nous retenons le premier, qui s'accorde avec « réflexion » comme « symétrie axiale » s'accorde avec « symétrie »), c'est-à-dire une réflexion suivie d'une translation parallèle à l'axe.

Nous utiliserons ce théorème sans le démontrer. Le tableau ci-dessous présente les quatre types d'isométries, en distinguant aussi l'identité.

Table 1.2 : Classification des isométries du plan suivant leur déterminant et l'ensemble de leurs points fixes.

Isométrie	$\det(A)$	Points fixes	Nom au secondaire
Identité	$+1$	tout le plan	—
Translation	$+1$	aucun	translation
Rotation	$+1$	un seul point	rotation
Réflexion	-1	une droite	réflexion, symétrie axiale
Réflexion glissée	-1	aucun	(absente du programme)

Pour la classe 1.7 **La réflexion glissée au-delà du programme**

La réflexion glissée ne figure pas au programme du secondaire québécois. On peut cependant la reconnaître dans les traces de pas dans la neige, la disposition des feuilles sur certaines tiges et plusieurs frises décoratives. Son existence montre aussi que la liste « translation, rotation, réflexion » est *incomplète*. En effet, la composée d'une réflexion et d'une translation peut n'appartenir à aucune de ces trois familles. Chercher à décrire une telle composée constitue une activité permettant de prolonger l'étude des isométries.



Figure 1.9 : Une frise engendrée par une réflexion glissée : chaque étape comprend une réflexion suivie d'une translation d'un demi-pas. Aucune translation, rotation ou réflexion seule ne produit ce motif.

1.8.3 La décomposition en réflexions

Théorème 1.13 **Composée de deux réflexions**

Soient d_1 et d_2 deux droites du plan et R_1, R_2 les réflexions correspondantes.

1. Si $d_1 \parallel d_2$, alors $R_2 \circ R_1$ est la translation de vecteur $2\mathbf{v}$, où $\mathbf{v}$ est le vecteur perpendiculaire menant de d_1 à d_2.
2. Si d_1 et d_2 se coupent en M en formant un angle φ, alors $R_2 \circ R_1$ est la rotation de centre M et d'angle 2φ.

Démonstration. Démontrons le second cas ; le premier se traite de façon analogue. Quitte à translater, supposons $M = \mathbf{0}$. Si d_1 fait un angle α avec l'axe des x et d_2 un angle $\alpha + \varphi$, alors, par (1.2),

$$R_{\alpha+\varphi} R_\alpha = \begin{pmatrix} \cos(2\alpha + 2\varphi) & \sin(2\alpha + 2\varphi) \\ \sin(2\alpha + 2\varphi) & -\cos(2\alpha + 2\varphi) \end{pmatrix} \begin{pmatrix} \cos 2\alpha & \sin 2\alpha \\ \sin 2\alpha & -\cos 2\alpha \end{pmatrix}.$$

En développant et en regroupant à l'aide des formules d'addition, on obtient

$$R_{\alpha+\varphi} R_\alpha = \begin{pmatrix} \cos 2\varphi & -\sin 2\varphi \\ \sin 2\varphi & \cos 2\varphi \end{pmatrix} = r_{2\varphi}.$$

□

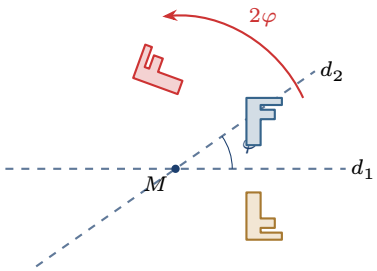


FIGURE 1.10 : Deux réflexions successives d'axes sécants donnent une rotation dont l'angle est le *double* de celui des axes. En bleu, le motif ; en or, son image par R_1 ; en rouge, l'image par $R_2 \circ R_1$.

Corollaire 1.14 **Trois réflexions suffisent**

Toute isométrie du plan est la composée d'au plus trois réflexions.

Pour la classe 1.8 **Le rôle de la réflexion dans l'enseignement**

Le corollaire 1.14 permet de mieux comprendre la place de la réflexion parmi les transformations du premier cycle du secondaire. En effet, toutes les isométries peuvent être obtenues en composant des réflexions ; c'est notamment le cas des translations et des rotations. Une activité avec deux miroirs ou deux pliages successifs permet d'observer ce résultat. Les élèves peuvent constater que deux réflexions d'axes sécants donnent une rotation et comparer l'angle de celle-ci avec l'angle entre les axes. La manipulation donne ainsi un appui concret à l'étude de cette propriété.

Outil 1.2 **Deux miroirs dans GeoGebra**

Le théorème 1.13 se prête bien à une exploration dans GeoGebra. On trace deux droites d_1 et d_2 et une figure F, puis on construit `F' = Reflect(F, d1)` et `F'' = Reflect(F', d2)`. En faisant tourner d_2 autour du point d'intersection, les élèves comparent l'angle `Angle(d1, d2)` avec l'angle de la rotation qui envoie F sur F'' ; en rendant les deux droites parallèles, ils voient la rotation devenir une translation. Ils peuvent aussi vérifier que l'ordre des réflexions compte : échanger d_1 et d_2 change le sens de la rotation.

1.9 Déterminer une transformation à partir de points

Jusqu'ici, nous avons cherché l'image d'une figure par une transformation donnée. Nous allons maintenant poser le problème dans l'autre sens : connaissant une figure et son image, comment déterminer la transformation qui les relie ? Cette question intervient notamment lorsqu'un logiciel redresse une photographie ou compare deux empreintes pour en reconnaître la forme.

Théorème 1.15 **Trois points suffisent**

Soient $P_1 = (x_1, y_1)$, $P_2 = (x_2, y_2)$ et $P_3 = (x_3, y_3)$ trois points *non alignés* du plan, et soient $Q_1 = (u_1, v_1)$, $Q_2 = (u_2, v_2)$, $Q_3 = (u_3, v_3)$ trois points quelconques. Alors il existe une *unique* application affine w telle que

$$w(P_1) = Q_1, \qquad w(P_2) = Q_2, \qquad w(P_3) = Q_3.$$

La démonstration consiste à traduire les conditions géométriques par un système d'équations linéaires. Cette méthode se prête directement à une résolution informatique.

Démonstration. Cherchons a, b, c, d, e, f tels que $w\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix}$. Les trois conditions $w(P_i) = Q_i$ s'écrivent

$$\begin{aligned} ax_1 + by_1 + e &= u_1, & cx_1 + dy_1 + f &= v_1, \\ ax_2 + by_2 + e &= u_2, & cx_2 + dy_2 + f &= v_2, \\ ax_3 + by_3 + e &= u_3, & cx_3 + dy_3 + f &= v_3. \end{aligned}$$

C'est un système de six équations à six inconnues. On peut le séparer en deux systèmes indépendants de trois équations à trois inconnues. De plus, ces deux systèmes possèdent la même matrice :

$$\begin{pmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{pmatrix}\begin{pmatrix} a \\ b \\ e \end{pmatrix} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \quad \text{et} \quad \begin{pmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{pmatrix}\begin{pmatrix} c \\ d \\ f \end{pmatrix} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}. \tag{1.5}$$

Notons M cette matrice commune. Un calcul direct donne

$$\det(M) = (x_2 - x_1)(y_3 - y_1) - (x_3 - x_1)(y_2 - y_1),$$

quantité qui vaut le double de l'aire orientée du triangle $P_1P_2P_3$: elle est non nulle exactement lorsque les trois points ne sont pas alignés. Dans ce cas, chacun des deux systèmes possède une solution unique, d'où l'existence et l'unicité de w. □

Remarque 1.10 **Une écriture plus compacte**

Les coordonnées homogènes permettent de résumer ce calcul. Posons

$$\widetilde{P} = \begin{pmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ 1 & 1 & 1 \end{pmatrix}, \qquad \widetilde{Q} = \begin{pmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ 1 & 1 & 1 \end{pmatrix}.$$

La matrice 3×3 de w s'obtient par

$$\widetilde{W} = \widetilde{Q}\,\widetilde{P}^{-1}.$$

Cette formule est celle qu'implémentent les bibliothèques d'infographie, et nous la retrouverons en Python à l'annexe A.

Exemple 1.10 **Le problème des deux triangles**

Cherchons l'application affine qui envoie le triangle de sommets $A = (-11, 3)$, $B = (1, 12)$, $C = (-1, 14)$ sur le triangle de sommets $P = (90, -68)$, $Q = (69, -53)$, $R = (83, -59)$, avec $A \mapsto P$, $B \mapsto Q$ et $C \mapsto R$.
Les images prescrites donnent les six équations suivantes :

$$\begin{aligned} -11a + 3b + e &= 90, & -11c + 3d + f &= -68, \\ a + 12b + e &= 69, & c + 12d + f &= -53, \\ -a + 14b + e &= 83, & -c + 14d + f &= -59. \end{aligned}$$

Le premier système donne $a = -4$, $b = 3$, $e = 37$; le second, $c = 2$, $d = -1$, $f = -43$. La transformation cherchée est donc

$$w\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -4 & 3 \\ 2 & -1 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 37 \\ -43 \end{pmatrix}.$$

Observons aussi que $\det(A) = 4 - 6 = -2$: la transformation renverse l'orientation et double les aires. Ce n'est ni une isométrie, ni même une similitude : une similitude linéaire de rapport k doit vérifier $A^\top A = k^2 I$, et $A^\top A = \begin{pmatrix} 20 & -14 \\ -14 & 10 \end{pmatrix}$ n'est pas un multiple de I.

Outil 1.3 **Résoudre les systèmes avec WolframAlpha**

Les deux systèmes de l'exemple 1.10 se résolvent en une ligne dans WolframAlpha : `solve -11a+3b+e=90, a+12b+e=69, -a+14b+e=83`. Le calcul étant confié à l'outil, l'attention se porte sur la *mise en équations*, qui est l'étape où se trouve la géométrie : chaque sommet et son image fournissent deux équations, et il faut trois couples de points pour obtenir les six équations nécessaires. On peut ensuite demander aux élèves de contrôler la réponse en calculant à la main l'image d'un des sommets.

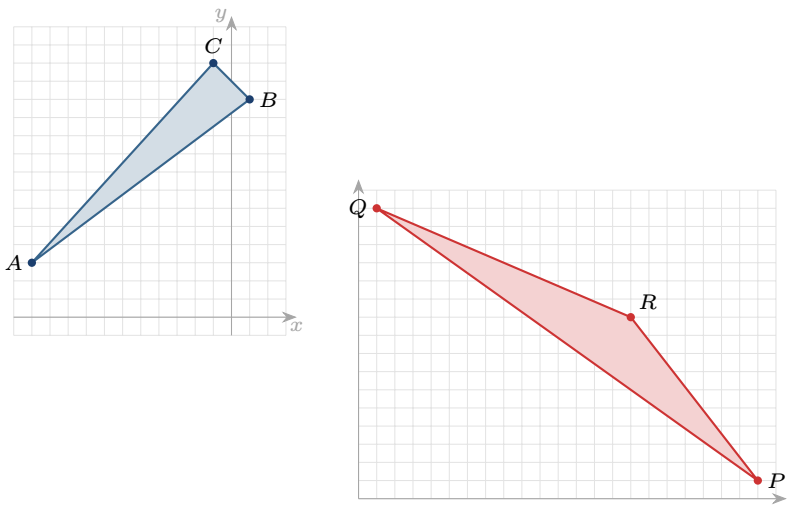


FIGURE 1.11 : Détermination de l'unique application affine qui envoie le triangle de gauche sur celui de droite, avec $A \mapsto P$, $B \mapsto Q$ et $C \mapsto R$. Le repère de droite est décalé pour tenir sur la page ; les deux quadrillages ont la même maille, ce qui permet de comparer les aires : celle de droite est bien le double de celle de gauche, conformément à $|\det(A)| = 2$.

Remarque 1.11 **Combien de points pour quel groupe ?**

Le théorème 1.15 a des analogues pour les sous-groupes. Une *similitude* du plan est entièrement déterminée par l'image de *deux* points distincts, et une *isométrie* aussi — à ceci près qu'il reste dans les deux cas une ambiguïté : deux solutions, l'une directe et l'autre indirecte, symétriques par rapport à la droite joignant les deux images. C'est exactement ce que dit l'exercice 1.10. Une *translation* est déterminée par l'image d'un seul point. Le nombre de points requis donne ainsi une façon concrète de comparer les contraintes qui caractérisent ces familles. Cette observation peut aider les élèves à distinguer, par exemple, les données nécessaires pour déterminer une rotation de celles qui définissent une translation.

1.10 Les transformations dans l'enseignement secondaire québécois

Nous allons maintenant examiner les liens entre les résultats de ce chapitre et l'enseignement secondaire. Nous situerons d'abord les transformations dans le programme québécois, puis nous comparerons les notations des manuels avec la représentation matricielle. Nous verrons ensuite comment ce point de vue peut aider à expliquer certaines propriétés et à comprendre les difficultés des élèves.

1.10.1 Où sont-elles dans le programme ?

Les transformations géométriques sont reprises à plusieurs moments du parcours scolaire, avec un vocabulaire et des représentations adaptés aux connaissances des élèves [Min07] [2] [Min16].

Deux observations peuvent être faites à partir de ce tableau. D'abord, l'ordre d'introduction

2. Disponible en ligne : https://cdn-contenu.quebec.ca/cdn-contenu/education/pfeq/secondaire/PFEQ-domaines-apprentissage-deuxieme-cycle-secondaire.pdf (voir le chapitre 4, « Domaines d'apprentissage »).

Table 1.3 : Introduction des transformations géométriques dans le programme québécois. Les niveaux indiqués correspondent au moment où la notion est introduite ; elle est reprise et enrichie ensuite.

Niveau	Ce qui est en jeu
Primaire, 3e cycle	Réflexion et axe de symétrie ; frises et dallages ; observation de régularités. Le vocabulaire est informel, le travail se fait par pliage, miroir et papier quadrillé.
1re secondaire	Translation, rotation et réflexion ; construction de l'image d'une figure ; figures isométriques ; frises et dallages analysés en termes de transformations.
2e secondaire	Homothétie ; figures semblables ; rapport de similitude ; premières justifications à l'aide des propriétés des transformations.
3e secondaire	Conditions minimales d'isométrie et de similitude des triangles ; recherche de mesures manquantes ; rapports k, k^2, k^3 pour les longueurs, les aires et les volumes.
4e secondaire	Géométrie analytique : distance, point de partage, droites. L'outillage algébrique nécessaire à la représentation analytique des transformations se met en place.
5e secondaire (SN)	Vecteurs : opérations, combinaisons linéaires, produit scalaire. Le vecteur de translation peut alors être décrit à l'aide de ces opérations, au-delà de sa représentation par une flèche.

des transformations rappelle la hiérarchie de Klein (figure 1.3) : les isométries précèdent les similitudes, tandis que les transformations affines générales ne sont pas étudiées. Rien ne permet toutefois d'affirmer que ce rapprochement est délibéré. D'autre part, le lien entre les transformations et leur représentation analytique n'est pas nécessairement établi au secondaire. Les élèves peuvent donc connaître les quatre transformations, les coordonnées et les vecteurs sans savoir comment ces notions se rejoignent. Le travail de ce chapitre permet à l'enseignant de faire cette synthèse.

1.10.2 Traduire la notation du secondaire

Les manuels québécois utilisent une notation fonctionnelle compacte pour désigner les transformations. Cette notation désigne la transformation elle-même et se traduit directement dans le langage matriciel. Une différence de convention doit cependant être précisée : là où nous écrivons R_d pour la réflexion d'axe d, les manuels écrivent s_d, du mot « symétrie ». La colonne de gauche ci-dessous reproduit leur notation, celle de droite la nôtre.

Pour la classe 1.9 — L'intérêt de la notation fonctionnelle

> Il arrive qu'on reproche à cette notation d'être lourde. Elle présente cependant un avantage : elle traite la transformation comme un *objet* qu'on peut nommer, comparer et composer, plutôt que comme une action qu'on exécute. La recherche en didactique souligne la difficulté de ce passage — passer de la transformation-*procédé* à la transformation-*objet* — et la notation aide à le franchir [Edw97]. Son utilisation peut donc soutenir cet apprentissage.

1.10.3 Les apports du point de vue matriciel

Même si les matrices ne sont pas enseignées au secondaire, leur étude permet de mieux comprendre plusieurs propriétés qui y sont utilisées. Voici six liens à retenir pour l'ensei-

TABLE 1.4 : Correspondance entre les notations du secondaire et la représentation analytique des transformations.

Secondaire	Se lit	Représentation analytique
$t_{(a,b)}$	translation de (a, b)	$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} a \\ b \end{pmatrix}$
$r_{O,\theta}$	rotation de centre O, d'angle θ	$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto r_\theta \begin{pmatrix} x \\ y \end{pmatrix}$
$r_{M,\theta}$	rotation de centre M	$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto r_\theta \left(\begin{pmatrix} x \\ y \end{pmatrix} - \mathbf{m} \right) + \mathbf{m}$
s_d	réflexion d'axe d	$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto R_\theta \begin{pmatrix} x \\ y \end{pmatrix}$ si d passe par O
$h_{O,k}$	homothétie de centre O, rapport k	$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto k \begin{pmatrix} x \\ y \end{pmatrix}$
$h_{M,k}$	homothétie de centre M	$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto k \begin{pmatrix} x \\ y \end{pmatrix} + (1-k)\mathbf{m}$
$s_d \circ t_{(a,b)}$	« on translate, puis on réfléchit »	$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto R_\theta \left(\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} a \\ b \end{pmatrix} \right)$

gnement.

1. **Un cadre commun.** Translation, rotation, réflexion et homothétie peuvent être décrites par une même formule $w \begin{pmatrix} x \\ y \end{pmatrix} = A \begin{pmatrix} x \\ y \end{pmatrix} + \mathbf{t}$. Quand un élève demande « est-ce qu'il existe d'autres transformations ? », ce cadre permet d'identifier d'autres exemples.
2. **Pourquoi l'ordre compte.** « Pourquoi $r \circ R \neq R \circ r$? » Parce que la multiplication des matrices n'est pas commutative (exemple 1.4). Le calcul algébrique et l'observation géométrique permettent ainsi d'expliquer le même phénomène.
3. **D'où vient le k^2.** La règle « les aires sont multipliées par k^2 » peut être justifiée par le déterminant, ou, plus simplement, par le fait qu'une aire est un produit de deux longueurs (remarque 1.6).
4. **Pourquoi les conditions minimales fonctionnent.** Les critères d'isométrie et de similitude des triangles (CAC, ACA, CCC, AA…) peuvent sembler arbitraires lorsqu'on les présente sans explication. Le théorème 1.15 permet d'éclairer cette question : une application affine est déterminée par l'image de trois points, et une similitude par l'image de deux. C'est pour cela qu'un nombre fini d'égalités suffit.
5. **Le rôle central de la réflexion.** Le corollaire 1.14 indique que la réflexion permet d'engendrer toutes les isométries. Ce rôle justifie qu'on prenne le temps de l'étudier, notamment en utilisant des miroirs.
6. **La liste des trois transformations est incomplète.** La réflexion glissée existe (remarque 1.7), et un élève qui compose une réflexion et une translation peut la rencontrer. Il est utile de pouvoir reconnaître et expliquer ce cas.

1.10.4 Les difficultés d'apprentissage, et d'où elles viennent

Les recherches en didactique font ressortir plusieurs difficultés récurrentes dans l'apprentissage des transformations. Certaines ne proviennent pas d'erreurs de calcul, mais de la façon dont les élèves, et parfois les futurs enseignants, interprètent les objets étudiés.

1. **La transformation vue comme un mouvement de la figure.** L'élève pense qu'on déplace le triangle, pas qu'on transforme le plan. Cela peut expliquer qu'il

ne conçoive pas l'image d'un point qui n'appartient pas à la figure, et qu'il ne comprenne pas qu'une rotation ait un centre situé « ailleurs » [Yan11]. On peut alors demander régulièrement l'image de points isolés, et l'image du *quadrillage* lui-même.

2. **La translation confondue avec son vecteur, ou avec un déplacement quelconque.** Beaucoup d'élèves croient que « translater » signifie « bouger sans tourner », sans se soucier de la distance ni de la direction. Une activité possible consiste à faire construire la translation par les seules coordonnées.
3. **Le centre de rotation systématiquement placé sur la figure.** Lorsque le centre est extérieur à la figure ou situé à l'intérieur de celle-ci, la construction peut poser des difficultés [Hol03]. On peut s'appuyer sur la conjugaison de la proposition 1.5, expliquée géométriquement, « je vais au centre, je tourne, je reviens ».
4. **L'homothétie réduite à un agrandissement.** L'élève n'imagine pas $k < 0$, et confond souvent le rapport k avec « le nombre de fois plus grand » en termes d'aire. Il est utile de faire tracer des homothéties de rapport négatif, et faire comparer systématiquement les longueurs, les aires et les angles.
5. **« Semblable » compris comme « qui se ressemble ».** Le sens courant du mot peut masquer sa définition mathématique. On peut revenir à la définition 1.11 elle-même, énoncée avec des mots — *toutes* les distances multipliées par le *même* nombre.
6. **Le niveau de van Hiele.** Plus généralement, les élèves qui raisonnent au niveau visuel (« ça a l'air pareil ») ne peuvent pas produire de justification fondée sur les propriétés, même s'ils obtiennent les bonnes réponses [Usi82][3]. Il faut donc examiner la justification, même lorsque la réponse obtenue est correcte.

Pour la classe 1.10 Une question pour préciser la conception de l'élève

La question suivante permet de voir si l'élève considère la transformation comme un mouvement limité à la figure : *« Soit une rotation de* 90° *autour d'un point M situé à l'extérieur du triangle ABC. Le point M a-t-il une image ? Laquelle ? »* L'élève qui répond « M n'est pas dans le triangle, donc la question ne se pose pas » ne prend pas en compte le fait que la transformation est définie sur le plan entier. C'est la même difficulté que celle de la remarque 1.1.

1.10.5 Quelques activités pour la classe

1. **Deux miroirs.** Placer deux petits miroirs formant un angle φ et compter les images. On découvre expérimentalement le théorème 1.13 : la composée de deux réflexions est une rotation d'angle 2φ. Faire varier φ jusqu'à obtenir un nombre entier d'images offre un problème d'exploration qui sera approfondi dans le cours MAT-2903.
2. **La chasse aux symétries.** Photographier des logos, des carrelages, des flocons, et compter, pour chacun, ses rotations et ses réflexions. La question « combien de symétries a ce logo ? » demande d'identifier précisément les transformations qui conservent la figure.
3. **Le pantographe.** Un instrument articulé qui réalise matériellement une homothétie. Il donne une interprétation concrète de la proposition 1.5, et la question « où est le centre ? » devient concrète : c'est le point où l'instrument est fixé.
4. **Programmer une transformation.** Faire écrire, en Python ou dans un logiciel de

3. Disponible en ligne : `https://files.eric.ed.gov/fulltext/ED220288.pdf`.

géométrie dynamique, la fonction qui calcule l'image d'un point. Comme nous l'avons noté à l'annexe A, la programmation demande de préciser des éléments qui peuvent rester implicites dans une construction à la main — en particulier le fait que la transformation s'applique à *tous* les points.

5. **La géométrie dynamique.** Construire une figure et son image, puis *déplacer* un sommet de la figure de départ et observer. Ce qui reste invariant pendant le déplacement, c'est précisément ce que la transformation conserve. Cette activité permet de donner un sens visuel à la notion d'invariant.

Pour la classe 1.11 **L'utilisation des outils**

L'utilisation d'un logiciel de géométrie dynamique ou d'un programme ne suffit pas, à elle seule, à assurer la compréhension. On peut manipuler un curseur ou exécuter un calcul sans chercher à interpréter le résultat. Une démarche utile consiste à demander aux élèves de *prédire* ce qu'ils pensent obtenir avant d'utiliser l'outil. Ils peuvent ensuite comparer leur prévision avec le résultat et chercher à expliquer les différences. L'outil sert alors à examiner une conjecture formulée au préalable [Hol03].

Exercices

Exercice 1.1. Parmi les correspondances suivantes dans le plan cartésien, identifier celles qui sont des transformations géométriques du plan.

a) $f_1 : (x, y) \mapsto (x^5, y)$
b) $f_2 : (x, y) \mapsto (\tan x, y)$
c) $f_3 : (x, y) \mapsto (x^2 + y^2, y)$
d) $f_4 : (x, y) \mapsto (5x - 1, -2y + 3)$
e) $f_5 : (x, y) \mapsto (y, x)$
f) $f_6 : (x, y) \mapsto (-3y, 3x)$
g) $f_7 : (x, y) \mapsto (\ln(1 + x^2), y)$
h) $f_8 : (x, y) \mapsto (x, \arctan y)$
i) $f_9 : (x, y) \mapsto (x + y, x - y)$

Exercice 1.2. Vrai ou faux ? Justifier chaque réponse en utilisant la propriété appropriée de la définition 1.1, ou en donnant un contre-exemple.

a) Une transformation du plan peut envoyer deux points distincts sur un même point.
b) Il est possible qu'un point du plan ne soit l'image d'aucun point par une transformation donnée.
c) Si f et g sont deux transformations telles que $f \circ g = \mathrm{id}$, alors $g \circ f = \mathrm{id}$ aussi.
d) Trois points alignés ont nécessairement des images alignées par *toute* transformation du plan.
e) Si w est une transformation *affine*, trois points alignés ont nécessairement des images alignées.
f) Une transformation du plan qui fixe trois points non alignés donnés est nécessairement l'identité.

Exercice 1.3. Montrer que la composée de deux transformations du plan est une transformation du plan, en vérifiant les quatre conditions de la définition 1.1.

Exercice 1.4. Soit l'application affine $w\begin{pmatrix}x\\y\end{pmatrix} = \begin{pmatrix}1 & -1\\2 & 0\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix} + \begin{pmatrix}-1\\1\end{pmatrix}$.

a) Déterminer l'image du rectangle plein $[0,2]\times[0,1]$.

b) Déterminer l'image du triangle plein de sommets $A=(1,1)$, $B=(0,0)$ et $C=(-1,1)$.

c) Déduire des propriétés de la transformation le rapport entre l'aire de chaque figure et celle de son image.

Exercice 1.5. Soit w une application linéaire. Montrer que w est une similitude de rapport k si et seulement si $\|w(U)\| = k$ pour tout vecteur U tel que $\|U\| = 1$.

Exercice 1.6. Montrer que les translations, les homothéties, les rotations et les réflexions sont des similitudes, et donner le rapport de chacune.

Exercice 1.7. Déterminer l'image du point (x,y) par l'homothétie de centre (h,k) et de rapport c. *Suggestion* : translater d'abord selon $(-h,-k)$, effectuer l'homothétie, puis translater selon (h,k).

Exercice 1.8. Déterminer l'expression analytique de l'isométrie qui envoie le triangle de sommets $(0,0)$, $(1,0)$, $(0,\frac{1}{2})$ sur le triangle image dont l'hypoténuse fait un angle de $\pi/4$ avec l'axe des x, le sommet $(0,0)$ étant envoyé sur $(1-\frac{\sqrt{2}}{4},\frac{\sqrt{2}}{4})$ et l'angle de rotation valant $-3\pi/4$.

Exercice 1.9. Déterminer l'expression analytique d'une application affine qui envoie le triangle de sommets $A=(-11,3)$, $B=(1,12)$, $C=(-1,14)$ sur le triangle de sommets $P=(90,-68)$, $Q=(69,-53)$, $R=(83,-59)$, avec $A\mapsto P$, $B\mapsto Q$ et $C\mapsto R$. Utiliser un outil informatique pour faciliter la résolution des systèmes.

Exercice 1.10. Déterminer deux similitudes qui envoient le segment $\mathcal{S}$ reliant $P=(-3,-1)$ à $Q=(-1,1)$ sur le segment $\mathcal{T}$ reliant $A=(3,1)$ à $B=(2,2)$, avec $P\mapsto A$ et $Q\mapsto B$. Y en a-t-il d'autres ?

Exercice 1.11. Déterminer l'expression analytique de l'application affine correspondant à la projection orthogonale sur la droite d'équation $y=x$. Est-ce une transformation du plan ?

Exercice 1.12. Soient $a,b,r>0$ et $e,f\in\mathbb{R}$. Déterminer l'équation du lieu obtenu comme image du cercle $x^2+y^2=r^2$ par $w\begin{pmatrix}x\\y\end{pmatrix} = \begin{pmatrix}a & 0\\0 & b\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix} + \begin{pmatrix}e\\f\end{pmatrix}$?

Exercice 1.13. Existe-t-il une similitude envoyant le cercle $x^2+y^2=4$ sur l'ellipse $x^2+\frac{y^2}{4}=1$?

Exercice 1.14. On admet que l'aire de $w(D)$ vaut $|\det(A)|$ fois celle de D.

a) Donner un exemple de transformation affine qui réduit les aires ($|\det(A)|<1$) sans être une contraction.

b) Établir que cette situation ne peut pas se produire pour une similitude : si w est une similitude, alors $|\det(A)|<1$ si et seulement si w est une contraction.

Exercice 1.15. Soit le triangle rectangle $\mathcal{T} = \triangle BAC$, d'angle droit en A, dont les côtés mesurent $a = \overline{BC}$, $b = \overline{AC}$ et $c = \overline{AB}$. Soit D le pied de la hauteur issue de A ; le triangle $\mathcal{T}$ est la réunion des triangles rectangles $\mathcal{T}_1 = \triangle ADC$ et $\mathcal{T}_2 = \triangle BDA$.

a) Montrer que $\mathcal{T}_1$ et $\mathcal{T}_2$ sont semblables à $\mathcal{T}$.

b) Il existe donc des similitudes w_1 et w_2 telles que $w_1(\mathcal{T}) = \mathcal{T}_1$ et $w_2(\mathcal{T}) = \mathcal{T}_2$. Déterminer les rapports k_1 et k_2.

c) À l'aide de la proposition 1.11 et l'égalité $s = s_1 + s_2$ entre les aires, démontrer le théorème de Pythagore.

d) Trouver un ensemble $\mathcal{E}$ tel que $w_1(\mathcal{E}) \cup w_2(\mathcal{E}) = \mathcal{E}$.

Exercice 1.16. Soit R la réflexion d'axe Ox et t la translation de vecteur $(3, 0)$. Montrer que $t \circ R$ est une réflexion glissée, et préciser son axe et son vecteur de glissement. Est-ce encore le cas pour $t \circ R$ si le vecteur de translation est plutôt $(0, 3)$?

Exercice 1.17. Écrire une fonction Python `decompose(A)` qui, étant donné une matrice 2×2 qui est une similitude, renvoie le rapport k, l'angle θ et le genre (directe ou indirecte). Vérifier son fonctionnement sur les matrices du catalogue de la table 1.1.

Exercice 1.18. Un élève de deuxième secondaire affirme : « Si on applique une homothétie de rapport 3 à un rectangle, son aire est multipliée par 3. » Rédiger une réponse à cet élève. Elle doit (i) reconnaître ce qui est juste dans son raisonnement, (ii) comporter un contre-exemple numérique et (iii) proposer une explication qui ne fasse pas appel aux matrices.

Exercice 1.19. Au secondaire québécois, seule la séquence Technico-sciences utilise des matrices pour décrire des transformations [Min16] ; les élèves des séquences Culture, société et technique et Sciences naturelles les étudient sans elles (tableau 1.3).

a) Choisir deux résultats de ce chapitre qui peuvent éclairer l'enseignement des transformations au secondaire, même avec des élèves qui ne voient pas de matrices. Pour chacun, préciser le niveau visé et la forme, sans matrice, sous laquelle il pourrait être présenté aux élèves.

b) Pour l'un de ces deux résultats, décrire une difficulté d'apprentissage que le point de vue du chapitre permet à l'enseignant d'anticiper.

c) Selon vous, la programmation d'une transformation (en Python ou dans un logiciel de géométrie dynamique) a-t-elle sa place en classe ? À quelles conditions ?

2 Espaces métriques

La notion de distance est rencontrée dès l'école primaire, lorsqu'on mesure l'écart entre deux points, entre deux villes sur une carte ou entre deux nombres sur une droite graduée. Ces situations font appel à une même intuition : une petite distance correspond à des objets proches, et la distance ne dépend pas de l'ordre dans lequel on les nomme. Cependant, comment définir cette notion sur un ensemble dont les éléments ne sont pas des points du plan ? Peut-on, par exemple, parler d'une distance entre deux villes sur une sphère, entre deux mots de même longueur ou entre deux figures géométriques ?

Pour répondre à ces questions, nous allons dégager quelques propriétés de la distance euclidienne et les utiliser pour définir les **espaces métriques**. Une fonction qui satisfait ces propriétés sur un ensemble quelconque permet d'y introduire les notions de voisinage, de convergence et de continuité. Nous pourrons même définir une distance entre des *ensembles*, comme nous le ferons à la section 2.3.

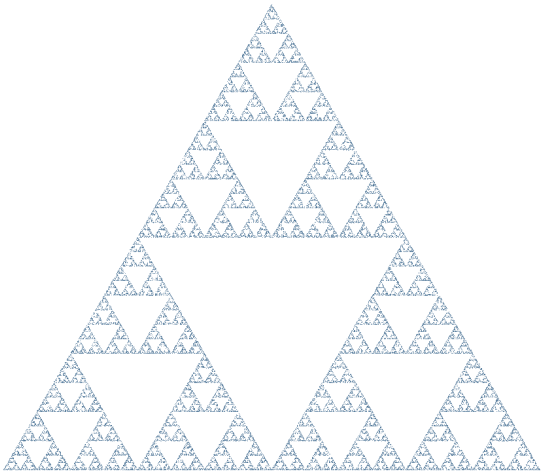


FIGURE 2.1 : Le *triangle de Sierpiński*, obtenu par un procédé aléatoire simple (le « jeu du chaos ») : partant d'un point, on choisit au hasard, à répétition, l'un des trois sommets d'un triangle, et on avance à mi-chemin vers lui. L'existence et l'unicité de l'ensemble fractal associé à cette construction seront justifiées au chapitre 3, à l'aide du théorème du point fixe de Banach (théorème 2.3, section 2.7), que nous démontrerons dans les pages qui suivent. Un tel ensemble fractal sera défini comme un point d'un espace dont les « points » sont eux-mêmes des ensembles compacts du plan, muni précisément de la distance de Hausdorff définie à la section 2.3. Cette application explique l'intérêt d'introduire une distance entre des ensembles.

Nous commencerons par la définition d'un espace métrique et par des exemples issus de la géométrie et de deux applications : la navigation sur une sphère, et la détection d'erreurs dans une chaîne de bits. Nous introduirons ensuite les notions d'ensemble ouvert, fermé

et compact, ce qui nous permettra de définir la distance de Hausdorff entre ensembles compacts. Nous terminerons par la notion de point fixe et par le théorème de Banach, qui établit, sous certaines hypothèses, l'existence et l'unicité d'un point fixe. Ce résultat est utilisé notamment dans les méthodes numériques, l'analyse ou la géométrie des fractales, objet du chapitre 3.

2.1 Qu'est-ce qu'une distance ?

Commençons par rappeler les propriétés de la distance euclidienne dans le plan. Si P, Q et R sont trois points du plan, et si $d(P, Q)$ désigne la distance entre P et Q :

1. $d(P, Q)$ est toujours un nombre positif ou nul ;
2. $d(P, Q) = 0$ exactement quand P et Q sont le même point ;
3. $d(P, Q) = d(Q, P)$, c'est-à-dire que la distance de P à Q est la même que celle de Q à P ;
4. $d(P, Q) \leq d(P, R) + d(R, Q)$, c'est-à-dire qu'un détour par un troisième point R n'est jamais plus court que le chemin direct.

Ces quatre propriétés ne font intervenir ni les coordonnées, ni le fait que P, Q et R soient des points du plan. Nous pouvons donc les reprendre pour définir une distance sur un ensemble quelconque.

Définition 2.1 **Métrique, espace métrique**

Soit E un ensemble. On appelle **métrique** (ou **distance**) sur E une fonction

$$d : E \times E \to \mathbb{R}$$

telle que, pour tout $x, y, z \in E$:

(a) $d(x, y) \geq 0$ (positivité)

(b) $d(x, y) = 0 \iff x = y$ (séparation)

(c) $d(x, y) = d(y, x)$ (symétrie)

(d) $d(x, y) \leq d(x, z) + d(z, y)$ (inégalité triangulaire)

Le couple (E, d) s'appelle alors un **espace métrique**.

L'expression « inégalité triangulaire » vient du cas $E = \mathbb{R}^2$. Dans un triangle de sommets x, y, z, la longueur de chaque côté est inférieure ou égale à la somme des longueurs des deux autres. La propriété étudiée au secondaire se retrouve donc ici dans une définition qui s'applique à des ensembles beaucoup plus généraux.

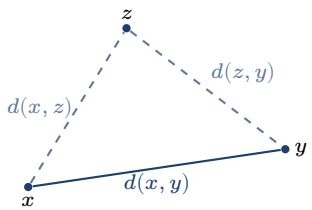


FIGURE 2.2 : L'inégalité triangulaire : la distance directe $d(x, y)$ est inférieure ou égale à la somme $d(x, z) + d(z, y)$, quel que soit le point z utilisé pour faire un détour.

Remarque 2.1 **La propriété de séparation**

La condition (b) réunit deux implications qu'il faut distinguer. D'une part, $d(x,x) = 0$ pour tout x : la distance d'un point à lui-même est nulle, conformément à l'intuition habituelle. D'autre part, *seuls* deux points confondus sont à distance nulle : si $d(x,y) = 0$, alors nécessairement $x = y$. Cette propriété porte le nom de **propriété de séparation**, parce qu'elle garantit qu'une métrique *sépare* toujours les points distincts par une distance strictement positive. Nous verrons à la section 2.3 un exemple où cette condition devra être vérifiée avec attention.

2.1.1 Les distances euclidienne et de Manhattan

Exemple 2.1 **La distance euclidienne dans $\mathbb{R}$ et $\mathbb{R}^n$**

Sur $\mathbb{R}$, la fonction $d_E(x,y) = |x-y|$ est une métrique : on vérifie les trois premières conditions à l'aide des propriétés élémentaires de la valeur absolue ; l'inégalité triangulaire découle de $|a+b| \leq |a| + |b|$, appliquée à $a = x - z$ et $b = z - y$. Plus généralement, si $X = (x_1, \dots, x_n)$ et $Y = (y_1, \dots, y_n)$ sont deux éléments de $\mathbb{R}^n$, la fonction

$$d_E(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

est une métrique sur $\mathbb{R}^n$, appelée **distance euclidienne**. Cette métrique correspond à la notion usuelle de distance : pour $n = 2$, on retrouve le théorème de Pythagore appliqué à la différence des coordonnées. Sauf mention contraire, c'est cette métrique que nous utiliserons sur $\mathbb{R}^n$.

Exemple 2.2 **La taxi-distance**

Notons $X = (x_1, y_1)$ et $Y = (x_2, y_2)$ deux points de $\mathbb{R}^2$. La fonction

$$d_T(X,Y) = |x_2 - x_1| + |y_2 - y_1|$$

définit une autre métrique sur $\mathbb{R}^2$ (on vérifie l'inégalité triangulaire en l'appliquant séparément à chaque coordonnée, comme à l'exemple 2.1). On l'appelle **taxi-distance**, ou **distance de Manhattan** : c'est la distance que devrait parcourir un taxi contraint de rouler sur un quadrillage de rues perpendiculaires, sans pouvoir emprunter de trajet diagonal.

Remarque 2.2 **La même paire de points, deux distances différentes**

La figure 2.3 permet de faire une distinction : d_E et d_T sont deux métriques *distinctes* définies sur le *même* ensemble $\mathbb{R}^2$. En général, $d_E(X,Y) \neq d_T(X,Y)$: dans l'exemple de la figure, $\sqrt{34} \approx 5{,}83$ contre 8. Un ensemble muni de deux métriques différentes donne lieu à deux espaces métriques différents, $(\mathbb{R}^2, d_E)$ et $(\mathbb{R}^2, d_T)$, bien qu'ils aient les mêmes éléments. Nous reverrons ce phénomène à la section 2.2 : la forme d'une boule ouverte dépend de la métrique choisie, pas seulement de l'ensemble sur lequel elle est définie.

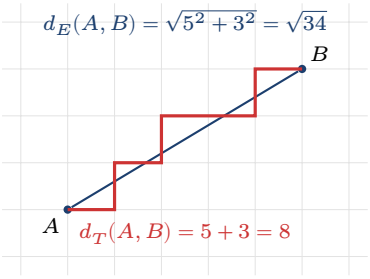


FIGURE 2.3 : Comparaison de deux distances entre les points A et B. Le segment bleu correspond à la distance euclidienne $d_E(A,B)$; l'escalier rouge, à la taxi-distance $d_T(A,B)$, soit la longueur totale des déplacements horizontaux et verticaux, quel que soit le chemin en escalier choisi entre A et B.

2.1.2 Une métrique appliquée : la distance sur la sphère

Les distances précédentes ont été définies dans le plan. Lorsqu'on s'intéresse à un déplacement sur une surface courbe, la distance mesurée dans l'espace ambiant ne correspond pas nécessairement au trajet possible. Par exemple, le segment qui joint Montréal à Paris passe à l'intérieur de la Terre ; un avion ne peut pas le suivre. Il faut donc considérer une distance mesurée *le long de la sphère*.

Pour deux points d'une sphère, la distance entre u et v se mesure le long de l'arc de **grand cercle** qui les joint, où le grand cercle est l'intersection de la sphère avec le plan passant par u, v et le centre O (figure 2.4) : cet arc décrit le plus court chemin *le long de la surface*. On appelle cette longueur d'arc la **distance sphérique**, ou distance du grand cercle, entre u et v. Une définition rigoureuse de cette distance, à l'aide des coordonnées de u et v, ainsi que les calculs qui l'accompagnent, relèvent d'une étude plus détaillée consacrée aux systèmes de positionnement et aux coordonnées sphériques.

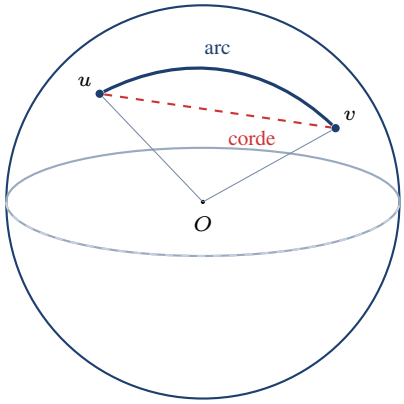


FIGURE 2.4 : La distance sphérique entre u et v se mesure le long de l'arc de grand cercle (en bleu), qui passe au-dessus de la corde (en rouge, pointillé) : cette corde traverse l'intérieur de la sphère, et ne représente donc pas un trajet sur sa surface. Cette distinction intervient dans la navigation aérienne et les systèmes GPS. Elle permet notamment de comprendre pourquoi les vols transatlantiques vers l'Europe empruntent une route qui semble courber vers le nord sur une carte plane.

2.1.3 Une métrique appliquée : mesurer les erreurs

Un autre exemple vient des communications numériques. Lorsqu'un message est transmis sous forme binaire, certains bits, soit des 0 ou des 1, peuvent être inversés sous l'effet du

bruit. Cela peut se produire dans un canal informatique, lors de la lecture d'un disque compact ou dans une transmission par satellite. Pour détecter ces erreurs et, si possible, les corriger, nous devons d'abord préciser comment *mesurer* la différence entre deux messages.

Définition 2.2 **Distance de Hamming**

Soit $n \in \mathbb{N}^\star$, et soit $\{0,1\}^n$ l'ensemble des suites de n bits. Pour $X = (x_1, \dots, x_n)$ et $Y = (y_1, \dots, y_n)$ dans $\{0,1\}^n$, on pose

$$d_{\text{Ham}}(X,Y) = \sum_{i=1}^{n} |x_i - y_i|.$$

On appelle d_{Ham} la **distance de Hamming** sur $\{0,1\}^n$.

Puisque $x_i, y_i \in \{0,1\}$, le terme $|x_i - y_i|$ vaut 1 si les deux bits diffèrent à la position i, et 0 sinon. Ainsi, $d_{\text{Ham}}(X,Y)$ donne ainsi le **nombre de positions où X et Y diffèrent**, c'est-à-dire le nombre minimal de bits qu'il faut inverser pour transformer X en Y.

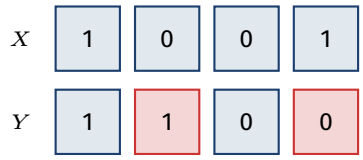
Exemple 2.3 **Un calcul de distance de Hamming**

Soit $X = (1,0,0,1)$ et $Y = (1,1,0,0)$ deux suites de $\{0,1\}^4$. Alors

$$d_{\text{Ham}}(X,Y) = |1-1| + |0-1| + |0-0| + |1-0| = 0 + 1 + 0 + 1 = 2.$$

Les suites X et Y diffèrent donc à exactement deux positions, la deuxième et la quatrième, comme on peut aussi le vérifier en comparant les termes :

$$X = (1, \underline{0}, 0, \underline{1}), \qquad Y = (1, \underline{1}, 0, \underline{0}).$$



Les trois premières conditions de la définition 2.1 se vérifient à partir de la somme : une somme de valeurs absolues est positive ou nulle, elle est nulle exactement quand chaque terme l'est (c'est-à-dire quand $X = Y$), et $|x_i - y_i| = |y_i - x_i|$ donne la symétrie. Pour l'inégalité triangulaire, il suffit d'observer que si X et Z diffèrent à la position i, ou si Z et Y diffèrent à la position i (ou les deux), alors X et Y peuvent différer à cette position, mais pas plus : $d_{\text{Ham}}(X,Y) \leq d_{\text{Ham}}(X,Z) + d_{\text{Ham}}(Z,Y)$, une position où X et Y diffèrent alors que X et Z n'y diffèrent pas forçant nécessairement Z et Y à y différer.

Note historique 2.1 **Richard Hamming et la détection d'erreurs**

La distance qui porte son nom a été introduite en 1950 par l'ingénieur et mathématicien états-unien Richard Hamming (1915–1998), qui travaillait alors aux laboratoires Bell sur la fiabilité des premiers ordinateurs à relais. Les arrêts de la machine durant les fins de semaine l'irritaient : celle-ci détectait une erreur de lecture, mais ne pouvait pas la corriger. Hamming s'est donc demandé comment construire un code qui permettrait aussi d'identifier *l'erreur* à corriger. Sa solution, les codes de Hamming, consiste à choisir les messages valides de façon qu'ils soient

tous à une distance de Hamming d'au moins 3 les uns des autres : un message corrompu par une seule erreur de bit reste alors strictement plus proche (au sens de d_{Ham}) du message original que de tout autre message valide, ce qui permet de la corriger automatiquement. On retrouve aujourd'hui cette même idée, sous une forme ou une autre, dans les CD, les disques durs et les communications par satellite.

2.2 Ouverts, fermés et compacts

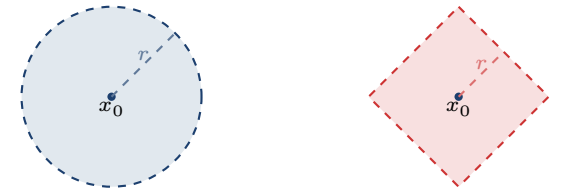
Avant de définir une distance entre deux *ensembles*, nous allons introduire les notions d'ensemble ouvert, fermé et compact. Elles s'appuient sur celle de boule ouverte et s'expriment dans le langage des espaces métriques. Nous les illustrerons principalement dans $\mathbb{R}^n$, où il est plus facile de leur donner une interprétation géométrique.

2.2.1 Boules et ensembles ouverts

Définition 2.3 **Boule ouverte**

Soit (E, d) un espace métrique, soit $x_0 \in E$ et soit $r > 0$. On appelle **boule ouverte** de centre x_0 et de rayon r l'ensemble

$$B(x_0, r) = \{x \in E : d(x, x_0) < r\}.$$



$B(x_0, r)$ dans $(\mathbb{R}^2, d_E)$ $B(x_0, r)$ dans $(\mathbb{R}^2, d_T)$

FIGURE 2.5 : L'allure d'une boule ouverte dépend de la métrique, même lorsque l'ensemble et le point x_0 sont les mêmes : un disque pour d_E (à gauche), un carré tourné de 45° pour d_T (à droite, car $|x - x_0| + |y - y_0| < r$ décrit précisément l'intérieur d'un tel carré). Comme nous l'avons annoncé à la remarque 2.2, la géométrie d'un espace métrique dépend autant de la métrique que de l'ensemble sous-jacent.

Outil 2.1 **Tracer des boules dans Desmos**

Desmos trace directement les régions définies par des inégalités. En entrant `x^2 + y^2 < r^2` et `|x| + |y| < r`, avec un curseur pour r, on obtient les deux boules de la figure 2.5 ; les bords sont automatiquement tracés en pointillé, ce qui rappelle qu'ils n'appartiennent pas à la boule ouverte. En remplaçant les inégalités strictes par `<=`, le bord devient plein : on dispose ainsi d'une représentation visuelle de la différence entre une boule ouverte et une boule fermée. Avec des élèves de quatrième secondaire, la même activité permet de comparer la « distance à vol d'oiseau » et la « distance en taxi » à partir d'un point donné.

Définition 2.4 **Ensemble ouvert**

Soit (E, d) un espace métrique, et soit $U \subseteq E$. On dira que U est un **ensemble ouvert** si, pour tout $x \in U$, il existe $r > 0$ tel que $B(x, r) \subseteq U$.

Ainsi, pour qu'un ensemble soit ouvert, chacun de ses points doit être le centre d'une boule entièrement contenue dans l'ensemble. Aucun point de U ne peut donc être sur sa frontière. Le rayon de la boule peut varier avec le point choisi : près de la frontière, il faudra généralement prendre un rayon plus petit, mais strictement positif.

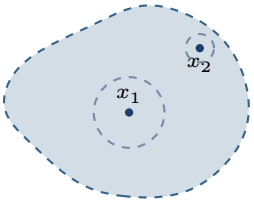


FIGURE 2.6 : Un ensemble ouvert (frontière en pointillé, pour rappeler qu'elle n'appartient pas à l'ensemble). Le point x_1, loin de la frontière, est le centre d'une boule de plus grand rayon entièrement contenue dans l'ensemble bleu ; le point x_2, plus près du bord, demande un rayon plus petit. Dans les deux cas, il existe une boule entièrement contenue dans l'ensemble, conformément à la définition 2.4.

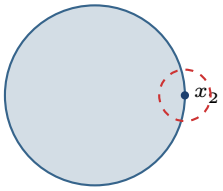


FIGURE 2.7 : Un ensemble qui n'est pas ouvert : le disque plein, bord compris. Le point x_2 appartient à la frontière incluse dans l'ensemble. Cependant, *toute* boule centrée en x_2, même minuscule (en pointillé), contient des points à l'extérieur du disque. Aucun rayon strictement positif ne satisfait donc la définition 2.4 en ce point.

La figure 2.7 illustre donc ce qui empêche un ensemble contenant un point de sa frontière d'être ouvert : aucune boule de rayon strictement positif centrée en ce point n'est entièrement contenue dans l'ensemble.

2.2.2 Ensembles fermés

Définition 2.5 **Ensemble fermé**

Soit (E, d) un espace métrique, et soit $F \subseteq E$. On dira que F est un **ensemble fermé** si son complément F^{c} est ouvert dans (E, d).

Remarque 2.3 **Ouvert, fermé : ni l'un ni l'autre, ou les deux**

Les mots « ouvert » et « fermé » n'ont pas ici le sens de deux possibilités qui s'excluent : un ensemble peut n'être ni l'un ni l'autre (pensons à un disque qui contient une partie, mais non la totalité, de sa frontière), et certains ensembles particuliers, comme E et l'ensemble vide $\varnothing$, sont à la fois ouverts et fermés dans (E, d).

2.2.3 Ensembles compacts

Nous aurons aussi besoin de la notion d'ensemble compact. Dans un espace métrique général, sa définition fait intervenir des recouvrements par des ouverts, notion que nous ne développerons pas ici. Le théorème suivant permet cependant de caractériser les compacts de $\mathbb{R}^n$ à l'aide de propriétés déjà rencontrées.

Définition 2.6 **Ensemble borné**

Soit (E, d) un espace métrique, et soit $F \subseteq E$. On dira que F est **borné** s'il existe un nombre réel c tel que $d(x, y) \leq c$ pour tout $x, y \in F$.

Théorème 2.1 **Heine-Borel**

Soit $K \subseteq \mathbb{R}^n$, muni de la distance euclidienne. L'ensemble K est **compact** si et seulement s'il est à la fois fermé et borné.

Dans la suite de ce chapitre, nous utiliserons donc « fermé et borné » comme définition d'un compact de $\mathbb{R}^n$. Cette convention est justifiée par le théorème 2.1 et permet de poursuivre sans développer la théorie des recouvrements ouverts.

Remarque 2.4 **La compacité dans un espace métrique général**

Précisons cependant que « fermé et borné » n'est pas, à proprement parler, la définition d'un ensemble compact : c'en est une *caractérisation*, valable dans $\mathbb{R}^n$ en vertu du théorème 2.1, mais qui cesse d'être vraie dans des espaces métriques plus généraux. La véritable définition, que nous ne détaillerons pas, exige que de *tout* recouvrement de K par des ensembles ouverts, on puisse extraire un nombre *fini* de ces ensembles qui recouvrent encore K, une propriété utilisée dans plusieurs résultats d'analyse. Pour nos besoins, dans $\mathbb{R}^n$, la caractérisation par « fermé et borné » sera suffisante.

2.3 La distance de Hausdorff

Notons $\mathcal{K}(E)$ l'ensemble des sous-ensembles compacts non vides d'un espace métrique (E, d). Nous souhaitons y définir une distance qui exprime l'idée suivante : deux ensembles sont proches lorsque chaque point de l'un est près d'un point de l'autre, et réciproquement. Pour préciser cette idée, nous allons d'abord définir un voisinage d'un ensemble. Au chapitre 3, cette construction permettra de considérer des figures fractales comme des points d'un espace métrique constitué de compacts du plan.

2.3.1 Le ε-voisinage et la somme de Minkowski

Définition 2.7 **Somme de Minkowski**

Soit $(E, d) = (\mathbb{R}^n, d_E)$, et soient $A, B \subseteq E$. On appelle **somme de Minkowski** de A et B l'ensemble

$$A \oplus B = \{a + b : a \in A,\ b \in B\}.$$

Autrement dit, $A \oplus B$ s'obtient en plaçant une copie de B à chaque point de A (ou, de façon équivalente, une copie de A à chaque point de B), puis en réunissant toutes ces copies. Lorsque B est une petite boule fermée centrée à l'origine, cette opération ajoute autour de A une région de même épaisseur dans toutes les directions.

Définition 2.8 **ε-voisinage d'un ensemble**

Soit $(E, d) = (\mathbb{R}^n, d_E)$, soit $A \subseteq E$ non vide, et soit $\varepsilon > 0$. On appelle **ε-voisinage** de A l'ensemble

$$A^{\oplus\varepsilon} = A \oplus \overline{B(0,\varepsilon)},$$

où $\overline{B(0,\varepsilon)} = \{x \in E : d_E(x,0) \leq \varepsilon\}$ est la boule *fermée* de rayon ε centrée à l'origine.

Nous pouvons aussi reformuler cette définition : $A^{\oplus\varepsilon}$ est l'ensemble des points de E situés à distance au plus ε d'*au moins un* point de A :

$$A^{\oplus\varepsilon} = \{x \in E : d_E(x,a) \leq \varepsilon \text{ pour un certain } a \in A\}.$$

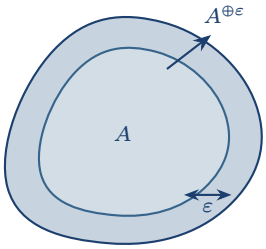


FIGURE 2.8 : Le ε-voisinage $A^{\oplus\varepsilon}$ (la région pleine, plus pâle, obtenue ici, pour simplifier le dessin, comme une homothétie agrandissant A plutôt que par la véritable somme de Minkowski) contient entièrement l'ensemble A (en bleu foncé, au centre) : $A \subseteq A^{\oplus\varepsilon}$, conformément à la définition 2.8. En effet, chaque point de A est à distance $0 \leq \varepsilon$ de lui-même.

Remarque 2.5 **La forme du voisinage dépend aussi de la métrique**

Comme les boules ouvertes (figure 2.5), le ε-voisinage d'un ensemble change d'aspect selon la métrique choisie. Pour $A = \{(0,0)\}$ un singleton, $A^{\oplus\varepsilon}$ est un disque plein de rayon ε si on somme avec une boule fermée de $(\mathbb{R}^2, d_E)$, mais un carré plein tourné de 45°, de « demi-diagonale » ε, si on utilise plutôt une boule fermée de $(\mathbb{R}^2, d_T)$: on retrouve donc la différence observée à la figure 2.5, dans le contexte du voisinage d'un ensemble.

2.3.2 Définition de la distance de Hausdorff

Définition 2.9 **Distance de Hausdorff**

Soit $(E, d) = (\mathbb{R}^n, d_E)$, et soient $A, B \in \mathcal{K}(E)$, c'est-à-dire deux sous-ensembles compacts non vides de E. On appelle **distance de Hausdorff** entre A et B le nombre

$$d_H(A,B) = \inf\{\varepsilon > 0 : A \subseteq B^{\oplus\varepsilon} \text{ et } B \subseteq A^{\oplus\varepsilon}\}.$$

Autrement dit, $d_H(A,B)$ est le plus petit rayon d'épaississement ε tel que chacun des deux ensembles soit entièrement recouvert par le ε-voisinage de l'autre. Puisque A et

B sont compacts, donc bornés, un tel ε existe toujours (on peut le choisir suffisamment grand), et on peut montrer que l'infimum est en fait atteint.

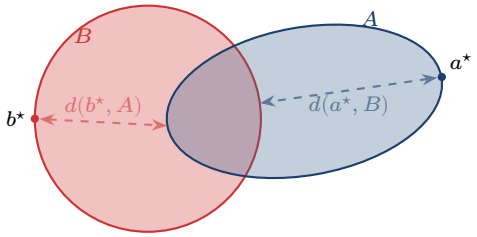


FIGURE 2.9 : Le point $a^\star \in A$ maximise la distance à B ; le point $b^\star \in B$ est celui qui est le plus loin de A. La distance de Hausdorff $d_H(A, B)$ est la plus grande de ces deux distances, $\max\{d(a^\star, B), d(b^\star, A)\}$, la formule (2.1) ci-dessous.

Il est souvent plus commode de calculer $d_H(A, B)$ à l'aide d'une formule équivalente, faisant intervenir la distance d'un point à un ensemble, $d(x, C) = \inf_{c\in C} d(x, c)$:

$$d_H(A, B) = \max\left\{\sup_{a\in A} d(a, B),\ \sup_{b\in B} d(b, A)\right\}. \tag{2.1}$$

Pour calculer $d_H(A, B)$, on cherche donc parmi les points de A celui dont la distance à B est la plus grande. On fait le même calcul en inversant les rôles des ensembles, puis on retient la plus grande des deux valeurs.

Exemple 2.4 **Deux carrés emboîtés**

Soit $A = [0, 2] \times [0, 2]$ le carré de côté 2, et soit $B = [0, 3] \times [0, 3]$ le carré de côté 3, dont le coin inférieur gauche est, dans les deux cas, l'origine. Tout point de A est à distance 0 de B (puisque $A \subseteq B$), donc $\sup_{a\in A} d(a, B) = 0$. Le point de B le plus éloigné de A est le coin opposé $(3, 3)$, à une distance de A égale à $d((3, 3), A) = d_E((3, 3), (2, 2)) = \sqrt{2}$. Ainsi, $d_H(A, B) = \max\{0, \sqrt{2}\} = \sqrt{2}$.

Remarque 2.6 **Pourquoi des ensembles compacts ?**

La restriction aux ensembles *compacts* dans la définition 2.9 permet notamment d'assurer la propriété de séparation, soit l'axiome (b) de la définition 2.1. Voyons ce qui peut arriver avec des sous-ensembles quelconques A et B de E. Prenons par exemple $A =]0, 1[$ et $B = [0, 1]$ dans $(\mathbb{R}, d_E)$: ces deux ensembles sont distincts, et pourtant $d_H(A, B) = 0$, car tout point de B est aussi près qu'on veut d'un point de A. Le problème vient de ce que A n'est pas fermé : il « manque » ses points frontières. En se restreignant aux ensembles compacts (donc fermés) de $\mathcal{K}(E)$, on peut démontrer que $d_H(A, B) = 0$ entraîne $A = B$, ce qui restaure la propriété de séparation annoncée à la remarque 2.1. On peut aussi vérifier que d_H satisfait la symétrie et l'inégalité triangulaire ; $(\mathcal{K}(E), d_H)$ est donc un espace métrique, dont les « points » sont eux-mêmes des ensembles.

Pour la classe 2.1 **Comparer deux figures géométriques**

La distance de Hausdorff permet de préciser une question qui peut survenir en classe : « à quel point ce triangle ressemble-t-il à celui-là ? » Deux polygones qui ont presque la même forme, mais dont un sommet a légèrement bougé, seront à une petite distance de Hausdorff l'un de l'autre ; des figures plus éloignées auront

une distance plus grande au sens de d_H. C'est d'ailleurs ce principe qu'utilisent certains logiciels de reconnaissance de formes, pour décider si un contour détecté dans une image ressemble suffisamment à un contour de référence.

Note historique 2.2 Felix Hausdorff



Felix Hausdorff (1868–1942). Domaine public, *Wikimedia Commons*.

Felix Hausdorff (1868–1942) était un mathématicien allemand, professeur aux universités de Leipzig, Bonn et Greifswald. Son ouvrage *Grundzüge der Mengenlehre* (« Fondements de la théorie des ensembles »), publié en 1914 [Hau14], a posé les bases de la topologie générale telle qu'on l'enseigne encore aujourd'hui ; c'est là qu'apparaît, entre autres, la distance qui porte son nom. Les intérêts de Hausdorff ne se limitaient pas aux mathématiques. Il a aussi publié, sous le pseudonyme littéraire *Paul Mongré*, des poèmes, des essais philosophiques sur Nietzsche et une pièce de théâtre satirique qui fut jouée avec succès de son vivant.

À partir de 1933, Hausdorff fut persécuté par le régime nazi en raison de ses origines juives : mis à la retraite forcée en 1935 « sans un mot de remerciement », il vit ses conditions se détériorer encore après le pogrom de la Nuit de Cristal en 1938. Une tentative d'émigration vers les États-Unis échoua en 1939. En janvier 1942, recevant l'ordre de se présenter à un camp d'internement, Hausdorff se donna la mort avec son épouse Charlotte et la sœur de celle-ci, plutôt que de s'y soumettre. Sa lettre d'adieu se terminait ainsi : « Ce qui s'est passé ces derniers mois contre les Juifs éveille la crainte justifiée qu'on ne nous laissera pas vivre jusqu'à une situation plus supportable. » Il avait 73 ans. Certains de ses travaux mathématiques, dont la notion de **dimension de Hausdorff** — un outil essentiel de la géométrie des fractales (remarque 3.3) —, n'a été pleinement reconnue qu'après sa mort [ORd].

2.4 Suites de Cauchy et complétude

Le théorème du point fixe de Banach, que nous démontrerons plus loin, établit l'existence d'une solution en construisant une suite qui converge vers elle. Il faut cependant s'assurer que la limite obtenue appartient à l'espace dans lequel on travaille. La notion de **complétude** permet d'exprimer cette exigence. Pour la définir, nous allons d'abord distinguer les suites convergentes des suites de Cauchy.

Définition 2.10 Suite convergente

Soit (E, d) un espace métrique, et soit (x_n) une suite d'éléments de E. On dira que (x_n) **converge** dans (E, d) s'il existe $\overline{x} \in E$ tel que, pour tout $\varepsilon > 0$, on peut trouver $N \in \mathbb{N}$ tel que $d(x_n, \overline{x}) < \varepsilon$ pour tout $n > N$. On dit alors que $\overline{x}$ est la **limite** de (x_n), et on note $\overline{x} = \lim_{n\to\infty} x_n$.

Examinons deux suites du plan afin d'illustrer la définition : la première converge, alors

que la seconde ne converge pas.

Exemple 2.5 **Une suite convergente dans le plan**

Soit $(x_n)_{n\geq 1}$ la suite de $\mathbb{R}^2$ définie par $x_n = (\frac{1}{n}, \frac{(-1)^n}{n})$. Les termes se trouvent alternativement de part et d'autre de l'axe des x et leur distance à l'origine diminue (figure 2.10) : pour $\varepsilon = 0{,}3$, on vérifie que $d_E(x_n, (0,0)) < 0{,}3$ dès que $n \geq 5$, de sorte que $N = 4$ convient. Plus généralement, $d_E(x_n, (0,0)) = \sqrt{2}/n \to 0$, donc (x_n) converge vers $\overline{x} = (0,0)$ au sens de la définition 2.10.

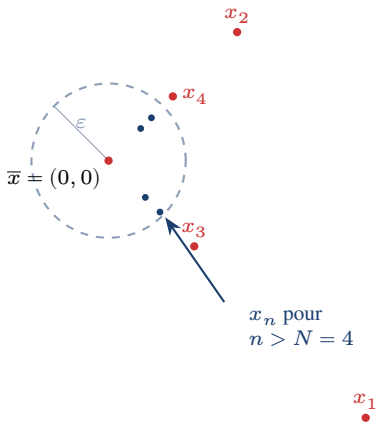


FIGURE 2.10 : Les quatre premiers termes $x_1, ..., x_4$ (en rouge) sont à l'extérieur du disque de rayon $\varepsilon = 0{,}3$ centré en $\overline{x}$; tous les termes suivants, $x_5, x_6, ...$ (en bleu), y sont contenus. Le rang $N = 4$ convient donc pour cette valeur de ε : on retrouve ainsi la condition de la définition 2.10.

Exemple 2.6 **Une suite divergente dans le plan**

Soit $(x_n)_{n\geq 1}$ la suite de $\mathbb{R}^2$ définie par $x_n = ((-1)^n, 0)$: elle vaut alternativement $(-1,0)$ et $(1,0)$, en alternant indéfiniment entre ces deux points (figure 2.11). Cette suite ne converge vers aucun point de $\mathbb{R}^2$: pour converger vers un point $\overline{x}$, il faudrait que $d_E(x_n, \overline{x})$ devienne aussi petit qu'on veut, ce qui est impossible tant que x_n continue d'osciller entre deux points distincts, à distance 2 l'un de l'autre.

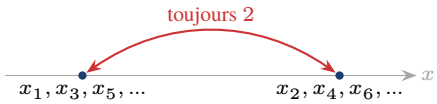


FIGURE 2.11 : La suite $x_n = ((-1)^n, 0)$ prend alternativement les valeurs $(-1,0)$ et $(1,0)$, toujours séparés d'une distance de 2 : elle ne converge vers aucun point de $\mathbb{R}^2$.

Définition 2.11 **Suite de Cauchy**

Soit (E,d) un espace métrique, et soit (x_n) une suite d'éléments de E. On dira que (x_n) est une **suite de Cauchy** dans (E,d) si, pour tout $\varepsilon > 0$, on peut trouver $N \in \mathbb{N}$ tel que $d(x_m, x_n) < \varepsilon$ pour tout $m, n > N$.

Remarque 2.7 **Les deux exemples et la condition de Cauchy**

La suite de l'exemple 2.5 est de Cauchy : puisqu'elle converge, la proposition 2.2 (énoncée ci-dessous) permet de le conclure. On peut aussi l'observer sur la figure 2.10 elle-même. Pour $m, n > N = 4$, les points x_m et x_n sont tous deux à l'intérieur du disque de rayon $\varepsilon = 0{,}3$, donc à distance au plus $2\varepsilon = 0{,}6$ l'un de l'autre : le même rang N qui fonctionne pour la convergence fonctionne, quitte à doubler ε, pour le critère de Cauchy. La suite de l'exemple 2.6, elle, n'est pas de Cauchy : quel que soit le rang N choisi, il existe toujours $m, n > N$ (un pair, un impair) tels que $d(x_m, x_n) = 2$. Aucun $\varepsilon < 2$ ne peut donc satisfaire la définition 2.11, peu importe N : ce qui est compatible avec la divergence déjà établie.

La distinction entre les deux définitions porte sur les distances que l'on compare. Pour une suite convergente, on mesure la distance des termes à un point donné, $\overline{x}$. Pour une suite de Cauchy, on compare les termes entre eux, sans avoir à identifier une limite. La proposition suivante établit le lien entre ces deux conditions.

Proposition 2.2

Soit (x_n) une suite convergente dans un espace métrique (E, d). Alors (x_n) est une suite de Cauchy dans (E, d).

Démonstration. Soit $\overline{x} = \lim_{n\to\infty} x_n$. Pour $\varepsilon > 0$ donné, il existe $N \in \mathbb{N}$ tel que $d(x_n, \overline{x}) < \varepsilon/2$ dès que $n > N$. Par l'inégalité triangulaire, si $m, n > N$,

$$d(x_m, x_n) \leq d(x_m, \overline{x}) + d(\overline{x}, x_n) < \varepsilon/2 + \varepsilon/2 = \varepsilon,$$

La condition de Cauchy est donc satisfaite : (x_n) est de Cauchy. □

Une suite de Cauchy n'est cependant pas toujours convergente. La notion de complétude permet de distinguer les espaces dans lesquels la réciproque est vraie.

Définition 2.12 **Espace métrique complet**

Soit (E, d) un espace métrique. On dit que (E, d) est **complet** si toute suite de Cauchy dans (E, d) converge vers un élément de E.

Exemple 2.7 **Les rationnels ne forment pas un espace complet**

Considérons $\mathbb{Q}$, l'ensemble des nombres rationnels, muni de la distance euclidienne $d_E(x, y) = |x - y|$. Soit (x_n) la suite des troncatures décimales de $\sqrt{2}$:

$$x_0 = 1, \quad x_1 = 1{,}4, \quad x_2 = 1{,}41, \quad x_3 = 1{,}414, \quad x_4 = 1{,}4142, \quad \ldots$$

c'est-à-dire que x_n conserve les n premières décimales de $\sqrt{2} = 1{,}414\,213\,56\ldots$ et tronque le reste. Chaque x_n est un nombre rationnel (une fraction décimale), et par construction $0 \leq \sqrt{2} - x_n < 10^{-n}$, de sorte que (x_n) converge vers $\sqrt{2}$ dans $(\mathbb{R}, d_E)$. Étant convergente dans $\mathbb{R}$, la proposition 2.2 nous assure que (x_n) est une suite de Cauchy. Mais $\sqrt{2} \notin \mathbb{Q}$: vue comme une suite de l'espace métrique $(\mathbb{Q}, d_E)$, elle est de Cauchy sans converger vers un élément de $\mathbb{Q}$. L'espace $(\mathbb{Q}, d_E)$ n'est donc *pas* complet : la limite de (x_n) dans les réels n'appartient pas à l'espace considéré.

Remarque 2.8 **La complétude des réels**

L'exemple 2.7 permet d'interpréter la complétude : $\mathbb{R}$ « complète » $\mathbb{Q}$ en ajoutant tous les nombres, comme $\sqrt{2}$ ou π, qui sont des limites de suites de rationnels sans être eux-mêmes rationnels. On admettra, sans le démontrer, le fait suivant : $(\mathbb{R}^n, d_E)$ est un espace métrique complet, quel que soit n. Cette propriété permettra d'appliquer le théorème de la section 2.7 dans plusieurs des exemples étudiés.

2.5 La notion de point fixe

Définition 2.13 **Point fixe**

Soit E un ensemble, et soit $f : E \to E$ une fonction. On dit que $x \in E$ est un **point fixe** de f si $f(x) = x$.

Un point fixe est un élément que la fonction envoie sur lui-même. Lorsque $E = \mathbb{R}$, on peut le repérer graphiquement à l'intersection du graphe de f et de la droite $y = x$ (figure 2.12). La définition conduit à deux questions : une fonction donnée possède-t-elle un point fixe et, si c'est le cas, est-il unique ? Les exemples suivants montrent que les réponses dépendent de la fonction considérée.

Exemple 2.8 **Trois fonctions de $\mathbb{R}$ dans $\mathbb{R}$**

(i) La fonction $f(x) = x + 1$ n'a *aucun* point fixe : l'équation $x + 1 = x$ n'a pas de solution.

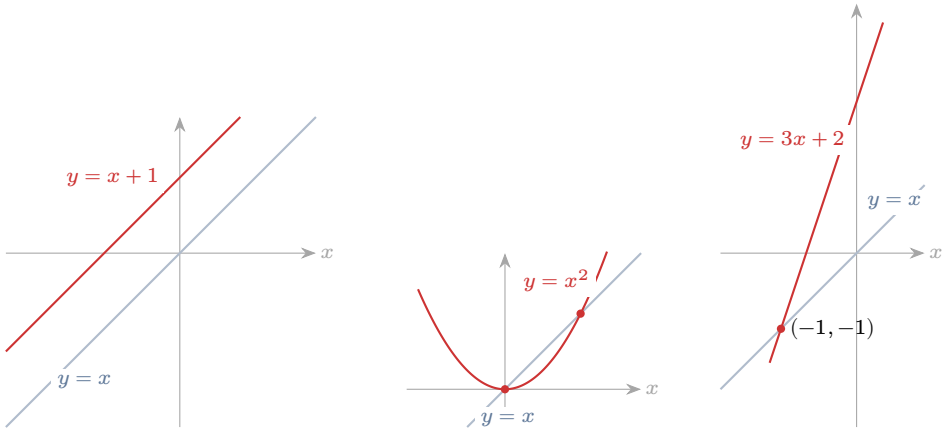
(ii) La fonction $f(x) = x^2$ a *exactement deux* points fixes : $x = 0$ et $x = 1$, les seules solutions de $x^2 = x$.

(iii) La fonction $f(x) = 3x + 2$ a *exactement un* point fixe, $x = -1$, seule solution de $3x + 2 = x$. Cependant, si on itère f à partir de n'importe quel $x_0 \neq -1$, la suite (x_n) s'*éloigne* de plus en plus de -1 plutôt que de s'en rapprocher : posséder un point fixe ne garantit donc pas que l'itération y conduise. Nous reverrons cette même fonction à la section 2.6, pour expliquer ce comportement.

Outil 2.2 **Points fixes et itérations dans Desmos et GeoGebra**

Dans Desmos, on trace $y = ax + b$ avec deux curseurs, ainsi que la droite $y = x$: le point fixe se lit à l'intersection, et les élèves peuvent conjecturer qu'il vaut $b/(1-a)$, puis remarquer qu'il disparaît lorsque $a = 1$ (le cas (i) de l'exemple 2.8). Dans GeoGebra, la commande `IterationList(f, x0, 20)` produit la liste $x_0, x_1, \ldots, x_{20}$ des itérés de f. En comparant $f(x) = 3x + 2$ et $f(x) = \frac{1}{3}x + 2$, on voit la suite s'éloigner du point fixe dans le premier cas et s'en approcher dans le second, ce qui prépare la notion de contraction (section 2.6).

La définition d'un point fixe n'est pas limitée aux fonctions réelles. Elle s'applique aussi aux fonctions sur $E = \mathbb{R}^2$ et sur tout autre ensemble.



(i) aucun point fixe (ii) deux points fixes (iii) un point fixe

FIGURE 2.12 : Les points fixes de f se repèrent aux intersections de son graphe avec la droite $y = x$ (en gris). (i) Les droites $y = x + 1$ et $y = x$, parallèles, ne se croisent jamais. (ii) La parabole $y = x^2$ croise $y = x$ exactement deux fois. (iii) La droite $y = 3x + 2$ la croise une seule fois, en $(-1, -1)$.

Exemple 2.9 **Une fonction de $\mathbb{R}^2$ dans $\mathbb{R}^2$**

Soit $f : \mathbb{R}^2 \to \mathbb{R}^2$ la fonction

$$f\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0.4 & -0.35 \\ 0.35 & 0.4 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix},$$

une similitude indirecte[a] de rapport $k \approx 0{,}531$. L'unique point fixe de f est l'origine $(0, 0)$: c'est le seul point qui soit envoyé sur lui-même par cette combinaison d'une rotation et d'une réduction. La figure 2.13 montre plusieurs images successives d'un même motif par f, qui s'enroulent en spirale vers ce point fixe.

a. On vérifie que la matrice a la forme $\begin{pmatrix} a & -b \\ b & a \end{pmatrix}$ associée à une similitude directe (section 1.8 du chapitre 1), de rapport $k = \sqrt{a^2 + b^2} = \sqrt{0.4^2 + 0.35^2} \approx 0{,}531$.

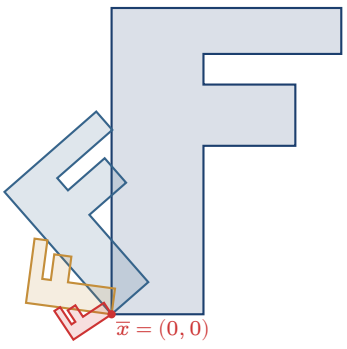


FIGURE 2.13 : Le motif « F » (déjà rencontré au chapitre 1), et ses images successives par la similitude f de l'exemple 2.9. Chaque nouvelle image est à la fois tournée et réduite d'un facteur $k \approx 0{,}531$; les images se rapprochent en spirale de l'unique point fixe, l'origine.

Remarque 2.9 **Un autre résultat : le théorème de Brouwer**

Le théorème vers lequel nous nous dirigeons (théorème 2.3) est l'un des résultats qui établissent l'existence d'un point fixe. Le théorème du point fixe de Brouwer, par exemple, affirme que toute fonction *continue* d'un disque fermé de $\mathbb{R}^2$ (ou, plus généralement, d'un ensemble convexe compact de $\mathbb{R}^n$) vers lui-même possède *au moins* un point fixe, sans aucune hypothèse de contraction. Ce théorème ne garantit cependant ni l'unicité (pensons à $f = \mathrm{id}$, dont tous les points sont fixes), ni une façon de *calculer* ce point fixe : il affirme son existence sans fournir de suite qui y converge. Avec des hypothèses plus restrictives, le théorème de Banach établit à la fois l'existence et l'unicité du point fixe. Il fournit aussi un algorithme pour le calculer et une estimation de la vitesse de convergence.

2.6 Contractions

Les exemples précédents invitent à examiner les fonctions qui rapprochent les points. Nous allons préciser cette propriété à l'aide de la notion de contraction, puis voir sous quelles conditions elle assure l'existence d'un unique point fixe.

Définition 2.14 **Contraction**

Soit (E, d) un espace métrique, et soit $f : E \to E$ une fonction. On dit que f est une **contraction** sur E s'il existe une constante $k \in [0, 1[$ telle que

$$d(f(x), f(y)) \leq k \cdot d(x, y) \qquad \forall\, x, y \in E.$$

La constante k s'appelle alors le **rapport de contraction** de f.

Exemple 2.10 **Contractions du plan**

Une homothétie de rapport k vérifiant $|k| < 1$ est une contraction de rapport $|k|$ sur $(\mathbb{R}^2, d_E)$; plus généralement, toute similitude de rapport $k < 1$, comme celle de l'exemple 2.9, en est une aussi. Une contraction n'est toutefois pas nécessairement une similitude : la fonction

$$f\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1/2 & 0 \\ 0 & 2/3 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}$$

est une contraction de rapport $2/3$ sur $(\mathbb{R}^2, d_E)$ (elle réduit davantage selon x que selon y), sans pourtant préserver les angles.

Remarque 2.10 **Le rôle de la borne stricte $k < 1$**

La condition $k < 1$, et non $k \leq 1$, est essentielle. La translation $f(x) = x + 1$ sur $(\mathbb{R}, d_E)$ vérifie $d(f(x), f(y)) = |x - y| = d(x, y)$, donc $1 \cdot d(x, y)$: elle conserve les distances, et nous avons vu à l'exemple 2.8 qu'elle n'a effectivement aucun point fixe. La fonction $f(x) = 3x + 2$, elle aussi de l'exemple 2.8, augmente pour sa part les distances : $d(f(x), f(y)) = 3|x - y| = 3\, d(x, y)$, donc non seulement $k = 3 \geq 1$, mais f *éloigne* systématiquement les points d'un facteur 3 au lieu de les rapprocher. Cela permet de comprendre que l'itération s'éloigne de son unique point fixe $x = -1$ plutôt que d'y converger.

2.7 Le théorème du point fixe de Banach

Les notions de complétude et de contraction permettent maintenant d'énoncer le théorème du point fixe de Banach.

Nous présenterons d'abord l'énoncé et sa démonstration dans $(\mathbb{R}^2, d_E)$, afin de pouvoir illustrer les étapes à l'aide de points du plan. Nous verrons ensuite que les arguments utilisés s'appliquent sans changement à un espace métrique complet (E, d) quelconque.

Théorème 2.3 **Théorème du point fixe de Banach**

Soit $f : \mathbb{R}^2 \to \mathbb{R}^2$ une contraction de rapport k (pour la distance d_E). Alors :

(i) f possède un unique point fixe $\overline{x} \in \mathbb{R}^2$;

(ii) pour tout $x_0 \in \mathbb{R}^2$, la suite (x_n) définie par $x_n = f^n(x_0)$ [a] converge vers $\overline{x}$, quel que soit le point de départ x_0 ;

(iii) on dispose de plus de la majoration

$$d_E(x_n, \overline{x}) \leq \frac{k^n}{1 - k}\, d_E(x_0, x_1).$$

a. La notation f^n désigne ici la composée de f avec elle-même n fois, $f^n = f \circ f \circ \cdots \circ f$, et non le produit de f par elle-même n fois.

Nous diviserons la démonstration en plusieurs lemmes. La démarche consiste à procéder de la façon suivante : on part d'un point x_0 *quelconque*, on construit la suite de ses images itérées $x_n = f^n(x_0)$, on montre qu'elle est de Cauchy (lemme 2.5), puis on invoque la complétude de $\mathbb{R}^2$ (remarque 2.8) pour affirmer qu'elle converge vers un certain $\overline{x} \in \mathbb{R}^2$. Nous vérifierons ensuite que cette limite est bien un point fixe (lemme 2.6), et qu'elle est unique (lemme 2.8).

Lemme 2.4 **Une contraction itérée reste une contraction**

Soit $f : \mathbb{R}^2 \to \mathbb{R}^2$ une contraction de rapport k. Alors, pour tout $n \in \mathbb{N}^\star$, f^n est une contraction de rapport k^n :

$$d_E(f^n(x), f^n(y)) \leq k^n\, d_E(x, y) \qquad \forall\, x, y \in \mathbb{R}^2\,.$$

Démonstration. Procédons par récurrence sur n. Lorsque $n = 1$, le résultat correspond à l'hypothèse. Supposons le résultat vrai pour un certain $n \geq 1$. Alors, en appliquant d'abord

que f est une contraction de rapport k (aux points $f^n(x)$ et $f^n(y)$), puis l'hypothèse de récurrence,

$$\begin{aligned} d_E(f^{n+1}(x), f^{n+1}(y)) &= d_E(f(f^n(x)), f(f^n(y))) \\ &\leq k\, d_E(f^n(x), f^n(y)) \\ &\leq k \cdot k^n\, d_E(x,y) = k^{n+1}\, d_E(x,y), \end{aligned}$$

ce qui établit le résultat pour $n+1$. □

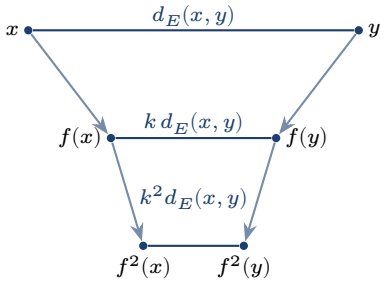


FIGURE 2.14 : Illustration du lemme 2.4 : à chaque application de f, la distance entre les images de x et y est multipliée par $k < 1$. Après n applications, elle vaut $k^n d_E(x,y)$, aussi petite qu'on veut dès que n est assez grand.

Lemme 2.5 **La suite des itérées est de Cauchy**

Soit $f : \mathbb{R}^2 \to \mathbb{R}^2$ une contraction de rapport k, soit $x_0 \in \mathbb{R}^2$, et soit $x_n = f^n(x_0)$. Alors, pour tout $n \in \mathbb{N}$ et tout $p \in \mathbb{N}$,

$$d_E(x_n, x_{n+p}) \leq \frac{k^n(1-k^p)}{1-k}\, d_E(x_0, x_1), \tag{2.2}$$

et par conséquent, (x_n) est une suite de Cauchy dans $(\mathbb{R}^2, d_E)$.

Démonstration. Par le lemme 2.4,

$$d_E(x_i, x_{i+1}) = d_E(f^i(x_0), f^i(x_1)) \leq k^i\, d_E(x_0, x_1)$$

pour tout $i \in \mathbb{N}$. En appliquant successivement p fois l'inégalité triangulaire, on relie x_n à x_{n+p} par une ligne brisée passant par tous les points intermédiaires (figure 2.15) :

$$\begin{aligned} d_E(x_n, x_{n+p}) &\leq d_E(x_n, x_{n+1}) + d_E(x_{n+1}, x_{n+2}) + \cdots + d_E(x_{n+p-1}, x_{n+p}) \\ &= \sum_{i=0}^{p-1} d_E(x_{n+i}, x_{n+i+1}) \leq \sum_{i=0}^{p-1} k^{n+i}\, d_E(x_0, x_1) \\ &= k^n\, d_E(x_0, x_1) \sum_{i=0}^{p-1} k^i = k^n\, d_E(x_0, x_1)\, \frac{1-k^p}{1-k}, \end{aligned}$$

ce qui établit (2.2).

Puisque $k \in [0, 1[$, on a $1 - k^p \leq 1$, d'où

$$d_E(x_n, x_{n+p}) \leq \frac{k^n}{1-k}\, d_E(x_0, x_1).$$

Puisque $0 \leq k < 1$, $k^n \to 0$ lorsque $n \to \infty$; le membre de droite tend donc vers 0, uniformément en p. Ainsi, pour tout $\varepsilon > 0$, on peut choisir N assez grand pour que

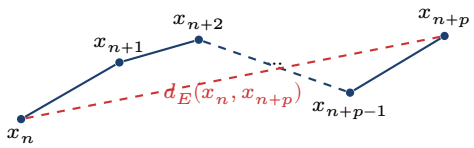


FIGURE 2.15 : La ligne brisée bleue joint x_n à x_{n+p} en passant par tous les termes intermédiaires de la suite ; sa longueur totale est la somme $\sum_i d_E(x_{n+i}, x_{n+i+1})$ qui apparaît dans le calcul ci-dessus. La distance directe $d_E(x_n, x_{n+p})$ (en rouge, pointillé) est la diagonale du polygone, éventuellement croisé, formé par ces points : par applications répétées de l'inégalité triangulaire, une diagonale n'excède jamais la somme des longueurs des côtés du polygone, c'est-à-dire son périmètre.

$d_E(x_n, x_{n+p}) < \varepsilon$ dès que $n > N$, quel que soit $p \geq 0$, ce qui revient à considérer $m, n > N$ en posant $m = n + p$. La suite (x_n) est donc de Cauchy. □

Puisque $(\mathbb{R}^2, d_E)$ est complet (remarque 2.8), le lemme 2.5 garantit l'existence d'un élément $\overline{x} \in \mathbb{R}^2$ tel que $\overline{x} = \lim_{n\to\infty} x_n$. Montrons maintenant que cette limite est un point fixe.

Lemme 2.6 **La limite est un point fixe**

Avec les notations précédentes, $\overline{x}$ est un point fixe de f, c'est-à-dire $f(\overline{x}) = \overline{x}$.

Démonstration. Une contraction est en particulier continue [1], donc elle commute avec la prise de limite. Ainsi,

$$f(\overline{x}) = f\Big(\lim_{n\to\infty} x_n\Big) = \lim_{n\to\infty} f(x_n) = \lim_{n\to\infty} x_{n+1} = \overline{x},$$

La dernière égalité vient du fait que (x_{n+1}) est la même suite que (x_n), à un décalage d'indice près, et possède donc la même limite. □

Proposition 2.7 **Majoration de l'erreur**

Avec les notations précédentes,

$$d_E(x_n, \overline{x}) \leq \frac{k^n}{1-k}\, d_E(x_0, x_1) \qquad \text{pour tout } n \in \mathbb{N}\,.$$

Démonstration. Faisons tendre $p \to \infty$ dans l'inégalité (2.2) du lemme 2.5, et utilisons de nouveau la continuité de $d_E(x_n, \cdot)$. Nous obtenons

$$d_E(x_n, \overline{x}) = \lim_{p\to\infty} d_E(x_n, x_{n+p}) \leq \lim_{p\to\infty} \frac{k^n(1-k^p)}{1-k}\, d_E(x_0, x_1) = \frac{k^n}{1-k}\, d_E(x_0, x_1),$$

puisque $k^p \to 0$ lorsque $p \to \infty$. □

Cette majoration permet de conclure à la convergence pour tout point de départ : puisque $k^n \to 0$, elle montre que $x_n \to \overline{x}$ *quel que soit* $x_0 \in \mathbb{R}^2$, la suite convergeant toujours vers le *même* point fixe, peu importe où on commence à itérer. Pour compléter la démonstration, établissons l'unicité du point fixe.

1. Si $d_E(x, y) < \delta$, alors $d_E(f(x), f(y)) \leq k\, d_E(x, y) < k\delta$, ce qui peut être rendu aussi petit qu'on veut : c'est exactement la continuité de f en tout point.

Lemme 2.8 **Unicité du point fixe**

Une contraction $f : \mathbb{R}^2 \to \mathbb{R}^2$ possède au plus un point fixe.

Démonstration. Supposons que f possède deux points fixes $\overline{x}_1$ et $\overline{x}_2$, avec $\overline{x}_1 \neq \overline{x}_2$. Puisque f est une contraction de rapport k,

$$d_E(\overline{x}_1, \overline{x}_2) = d_E(f(\overline{x}_1), f(\overline{x}_2)) \leq k\, d_E(\overline{x}_1, \overline{x}_2).$$

Puisque $\overline{x}_1 \neq \overline{x}_2$, la propriété de séparation donne $d_E(\overline{x}_1, \overline{x}_2) \neq 0$; on peut donc diviser les deux membres par cette quantité, on obtient alors $1 \leq k$, en contradiction avec l'hypothèse $k < 1$. Il ne peut donc pas exister deux points fixes distincts. □

En réunissant les lemmes 2.4, 2.5, 2.6 et 2.8, et la proposition 2.7, toutes les affirmations du théorème 2.3 sont établies.

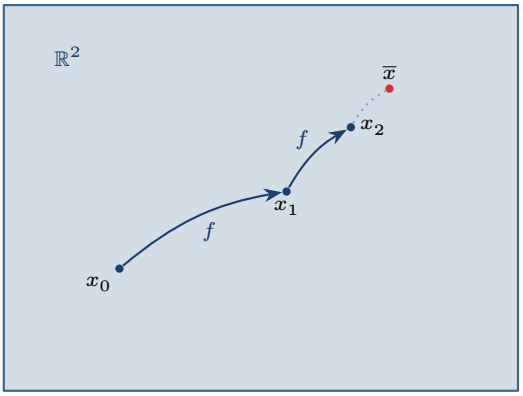


FIGURE 2.16 : Interprétation géométrique de la convergence : à chaque application de f, la distance au point fixe $\overline{x}$ est multipliée par un facteur $k < 1$ (définition 2.14), de sorte que les itérées $x_0, x_1 = f(x_0), x_2 = f(x_1), \ldots$ convergent vers $\overline{x}$, quel que soit le point de départ x_0.

Remarque 2.11 **Le cas d'un espace métrique quelconque**

Nous avons énoncé et démontré le théorème 2.3 dans $(\mathbb{R}^2, d_E)$ pour faciliter son illustration géométrique. Cependant, aucun argument dans les lemmes 2.4, 2.5, 2.6 et 2.8, ni dans la proposition 2.7, n'utilise une propriété particulière du plan : chaque démonstration ne se sert que des quatre axiomes d'une métrique (définition 2.1) et de la complétude de l'espace ambiant. En remplaçant partout $\mathbb{R}^2$ par un espace métrique complet (E, d) quelconque, et d_E par d, les mêmes démonstrations établissent le théorème dans toute sa généralité : si (E, d) est un espace métrique complet et $f : E \to E$ une contraction, alors f possède un unique point fixe $\overline{x}$, vers lequel converge la suite $(f^n(x_0))$ pour tout point de départ $x_0 \in E$, avec la même majoration géométrique de l'erreur. C'est sous cette forme générale que nous utiliserons le théorème dans le reste de ce chapitre, notamment à la section 2.7.1, où l'espace en jeu n'est pas $\mathbb{R}^2$ mais un intervalle fermé de $\mathbb{R}$. Nous l'utiliserons également au chapitre 3, où l'espace en jeu sera plutôt l'espace $(\mathcal{K}(\mathbb{R}^2), d_H)$ des ensembles compacts du plan que nous venons de construire.

2.7.1 La méthode de Newton-Raphson

La méthode de Newton, aussi appelée méthode de Newton-Raphson, fournit une application du théorème de Banach. Pour comprendre ce lien, nous allons reformuler la recherche d'un zéro comme un problème de point fixe.

Soit $f : \mathbb{R} \to \mathbb{R}$ une fonction deux fois dérivable, et soit $x^\star$ un zéro de f tel que $f'(x^\star) \neq 0$. Plutôt que de chercher directement à résoudre $f(x) = 0$, on définit une nouvelle fonction

$$g(x) = x - \frac{f(x)}{f'(x)},$$

partout où $f'(x) \neq 0$. Puisque $f(x^\star) = 0$, on a immédiatement $g(x^\star) = x^\star$: un zéro de f est un point fixe de g. La méthode consiste donc à itérer g plutôt que de résoudre $f(x) = 0$ directement.

Proposition 2.9 **La fonction g est localement une contraction**

Avec les notations précédentes, $g'(x^\star) = 0$. Par conséquent, pour tout $k \in]0, 1[$, il existe $\delta > 0$ tel que g soit une contraction de rapport k sur l'intervalle fermé $I = [x^\star - \delta, x^\star + \delta]$.

Démonstration. En dérivant, nous obtenons

$$g'(x) = 1 - \frac{f'(x)^2 - f(x)f''(x)}{f'(x)^2} = \frac{f(x)f''(x)}{f'(x)^2},$$

de sorte que $g'(x^\star) = 0$ puisque $f(x^\star) = 0$. Comme f est deux fois dérivable et que $f'(x^\star) \neq 0$, la fonction g' est continue au voisinage de $x^\star$; puisqu'elle s'y annule, on peut trouver $\delta > 0$ tel que $|g'(x)| \leq k$ pour tout $x \in I$, quitte à réduire δ pour s'assurer aussi que $g(I) \subseteq I$. Le théorème des accroissements finis donne alors, pour tous $x, y \in I$, $|g(x) - g(y)| = |g'(c)| \cdot |x - y| \leq k|x - y|$ pour un certain c entre x et y : g est bien une contraction de rapport k sur I. □

Remarque 2.12 **Vérification de la complétude de l'intervalle**

Le théorème 2.3 a été énoncé pour $(\mathbb{R}^2, d_E)$, et I est un intervalle de $\mathbb{R}$, pas du plan : c'est ici qu'intervient la remarque 2.11, qui étend le théorème à tout espace métrique complet (E, d). Il suffit donc de vérifier que (I, d_E) est complet. On admet le fait suivant, que l'on peut démontrer à partir de la définition 2.12 : tout sous-ensemble *fermé* d'un espace métrique complet est lui-même complet. Puisque I est un intervalle fermé de $\mathbb{R}$, qui est complet (remarque 2.8), (I, d_E) est bien complet, et le théorème 2.3 s'applique.

Le théorème 2.3, appliqué à g sur I, garantit donc qu'à partir de tout point de départ $x_0 \in I$, l'itération de Newton $x_{n+1} = g(x_n)$ converge vers $x^\star$. On peut préciser la vitesse de cette convergence. Puisque $g'(x^\star) = 0$ exactement (et non seulement $|g'(x^\star)| \leq k$ pour un k fixé d'avance), un développement de Taylor plus fin montre que la convergence est en réalité **quadratique**, c'est-à-dire que le nombre de décimales exactes double approximativement à chaque itération, une fois qu'on est assez près de $x^\star$, ce qui est plus rapide que la convergence géométrique garantie par le théorème de Banach seul.

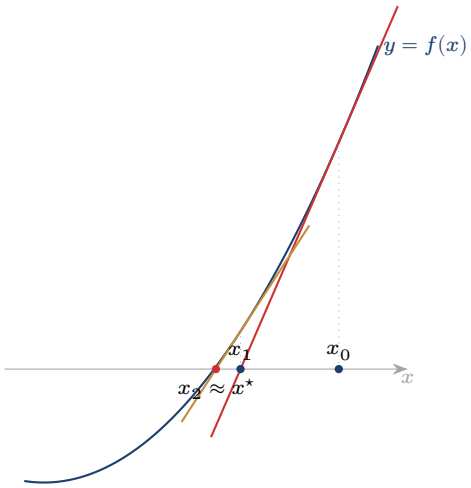


FIGURE 2.17 : L'interprétation géométrique de la méthode de Newton, pour $f(x) = (x-2)^2 - 3$: partant de $x_0 = 5$, on trace la tangente à la courbe en $(x_0, f(x_0))$ (en rouge) jusqu'à son intersection avec l'axe des x, ce qui donne $x_1 = 4$; on répète l'opération à partir de x_1 (tangente en or, cette fois en $(x_1, f(x_1))$) pour obtenir $x_2 = 3{,}75$, à comparer avec $x^\star = 2 + \sqrt{3} \approx 3{,}732$.

Exemple 2.11 **Deux méthodes, une même équation**

Considérons de nouveau l'équation $\cos(x) = x$ déjà rencontrée à la section 2.6, dont l'unique solution est $x^\star \approx 0{,}739\,085\,133\,2$. Nous savons déjà que la méthode simple $x_{n+1} = \cos(x_n)$ converge vers $x^\star$, puisque cos est une contraction de rapport $k = \sin(1) \approx 0{,}841$ sur $\mathbb{R}$ (remarque 2.10) ; partant de $x_0 = 0$, il faut 21 itérations pour obtenir 4 décimales exactes, et 45 pour en obtenir 8. Comparons avec la méthode de Newton appliquée à $f(x) = \cos(x) - x$, où $g(x) = x + \dfrac{\cos(x) - x}{\sin(x) + 1}$, partant cette fois de $x_0 = 1$:

n	x_n	erreur $\lvert x_n - x^\star \rvert$
0	1,000 000 000	$2{,}6 \times 10^{-1}$
1	0,750 363 868	$1{,}1 \times 10^{-2}$
2	0,739 112 891	$2{,}8 \times 10^{-5}$
3	0,739 085 133	$1{,}9 \times 10^{-10}$

Trois itérations de Newton suffisent donc pour dépasser la précision qu'il avait fallu 45 itérations de la méthode simple ($x_{n+1} = \cos(x_n)$, ci-dessus) pour atteindre : on observe ici la convergence quadratique, le nombre de décimales exactes doublant approximativement à chaque étape $(1, 2, 5, 10, \ldots)$.

Nous avons utilisé l'équation $\cos(x) = x$ pour comparer deux méthodes de calcul. Examinons maintenant un problème d'optimisation qui conduit à une équation polynomiale de degré 3 dont on ne reconnaît pas immédiatement de racine.

Exemple 2.12 **Une boîte à coût minimal**

Une entreprise fabrique des boîtes de carton à base carrée de côté x, sans couvercle, chacune devant contenir un volume fixé de $V = 800$ cm^3 ; la hauteur est donc $h = V/x^2$. Le carton de la base coûte $0{,}05$ \$/cm^2, celui des quatre parois latérales coûte $0{,}02$ \$/cm^2, et on ajoute, tout le long du pourtour de l'ouverture supérieure (un carré de périmètre $4x$), une bordure de renfort coûtant $0{,}08$ \$/cm. Le coût total d'une boîte est donc

$$C(x) = \underbrace{0{,}05\,x^2}_{\text{base}} + \underbrace{0{,}02 \cdot 4xh}_{\text{parois}} + \underbrace{0{,}08 \cdot 4x}_{\text{bordure}} = 0{,}05\,x^2 + \frac{64}{x} + 0{,}32\,x,$$

en substituant $h = 800/x^2$. Minimiser le coût revient à résoudre $C'(x) = 0$, où

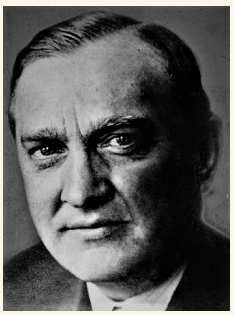
$$f(x) := C'(x) = 0{,}10\,x - \frac{64}{x^2} + 0{,}32.$$

En éliminant le dénominateur, $f(x) = 0$ équivaut à l'équation cubique $0{,}10\,x^3 + 0{,}32\,x^2 - 64 = 0$, dont on ne voit pas immédiatement de racine. Appliquons plutôt la méthode de Newton directement à f, avec $f'(x) = 0{,}10 + 128/x^3$, en partant de $x_0 = 8$ (une boîte presque cubique) :

n	x_n (cm)	$f(x_n)$
0	8,000 000	0,1200
1	7,657 143	$-5{,}84 \times 10^{-3}$
2	7,672 315	$-1{,}28 \times 10^{-5}$
3	7,672 348	$-6{,}20 \times 10^{-11}$

La boîte de coût minimal a donc un côté de base $x^\star \approx 7{,}672$ cm et une hauteur $h^\star = 800/(x^\star)^2 \approx 13{,}59$ cm, pour un coût total d'environ 13,74 \$. Trois itérations donnent ici une précision suffisante pour les dimensions d'une boîte fabriquée dans ce contexte.

Note historique 2.3 **Stefan Banach**



Stefan Banach (1892–1945). Domaine public, *Wikimedia Commons*.

Stefan Banach (1892–1945) est né à Cracovie, en Pologne, d'une mère trop pauvre pour l'élever et d'un père dont le statut militaire interdisait le mariage : il fut donc confié à des familles d'accueil. Il apprit ensuite une grande partie des mathématiques par lui-même. En 1916, le mathématicien Hugo Steinhaus, se promenant dans un parc de Cracovie, surprit une conversation entre Banach et un ami sur l'intégrale de Lebesgue qui était alors une notion récente. Intrigué, il lui proposa quelques problèmes ouverts. Banach les résolut en une semaine. Steinhaus racontera plus tard : « Ma plus grande découverte scientifique a été de découvrir Banach. »

Après être devenu professeur à Lwów (aujourd'hui Lviv, en Ukraine), Banach y fonda avec ses collègues l'*école mathématique de Lwów*, qui se réunissait au Café Écossais pour discuter de problèmes ouverts, consignés dans un cahier resté célèbre sous le nom de *Livre écossais*. Sa thèse de 1920, publiée en 1922 [Ban22], définit axiomatiquement ce qu'on appelle aujourd'hui un **espace de Banach** — un espace

vectoriel normé complet — et fonde ainsi l'analyse fonctionnelle moderne ; c'est dans ce même article qu'apparaît le théorème du point fixe qui porte son nom. Pendant l'occupation allemande de la Pologne, Banach échappa à la déportation en travaillant comme « nourricier de poux » dans un institut de recherche sur le typhus , emploi qui permettait alors de protéger de nombreux universitaires. Il mourut d'un cancer du poumon en 1945, peu après la fin de la guerre, à 53 ans [ORi].

Note historique 2.4 Les origines et les applications du théorème

L'idée d'obtenir une solution par approximations successives est plus ancienne que Banach : le mathématicien français Émile Picard l'utilisait déjà en 1890 pour démontrer l'existence et l'unicité des solutions de certaines équations différentielles [Pic90] — un résultat aujourd'hui connu sous le nom de théorème de Picard-Lindelöf. L'apport de Banach, en 1922, fut de présenter un cadre *abstrait* pour cette méthode. En isolant la complétude de l'espace et le caractère contractant de l'application, il a pu établir un résultat commun à plusieurs problèmes, sans reprendre la démonstration pour chaque équation. Le mathématicien italien Renato Caccioppoli redécouvrit indépendamment le même résultat en 1930 [Cac30], si bien qu'on le désigne parfois sous le nom de théorème de *Banach-Caccioppoli*.
Ce cadre général explique la diversité des applications du théorème : la construction d'ensembles fractals, que nous verrons au chapitre 3 ; la méthode de Newton et d'autres algorithmes numériques de résolution d'équations, comme à la section 2.7.1 ; ainsi que l'algorithme *PageRank*, qui a servi à classer les pages web chez Google à ses débuts — le calcul de l'importance relative des pages revient à trouver le point fixe d'une certaine contraction, dans un espace de dimension aussi grande que le nombre de pages du Web (chapitre 4).

Exercices

Exercice 2.1. Soit d la distance euclidienne sur $\mathbb{R}$. Montrer que la fonction

$$\delta(x, y) = \frac{|x - y|}{1 + |x - y|}$$

est aussi une métrique sur $\mathbb{R}$. (*Indication* : pour l'inégalité triangulaire, on pourra utiliser le fait que $t \mapsto t/(1 + t)$ est croissante sur $[0, \infty[$.)

Exercice 2.2. Un code correcteur d'erreurs utilise les quatre messages valides 0000, 0110, 1011 et 1101 dans $\{0, 1\}^4$.

a) Calculer la distance de Hamming entre chaque paire de messages valides.
b) Ce code peut-il corriger automatiquement toute erreur portant sur un seul bit ? Justifier à l'aide de la note historique 2.1.

Exercice 2.3. Pour chacun des sous-ensembles suivants de $(\mathbb{R}^2, d_E)$, indiquer lesquelles des propriétés « ouvert », « fermé » et « compact » sont satisfaites.

a) $A_1 = B((0, 0), 1)$ (une boule ouverte) ;
b) $A_2 = \{(x, y) : x^2 + y^2 \leq 1\}$;
c) $A_3 = \{(x, y) : x^2 + y^2 \leq 1\} \setminus \{(0, 0)\}$;

d) $A_4 = \mathbb{R}^2$.

Exercice 2.4. Soit $A = \{0\} \subseteq \mathbb{R}$ (un singleton) et $B = [-1, 1]$, tous deux munis de la distance euclidienne. Calculer $d_H(A, B)$.

Exercice 2.5. Soit $f : \mathbb{R} \to \mathbb{R}$ définie par $f(x) = \frac{1}{3}x + 2$.

a) Montrer que f est une contraction sur $(\mathbb{R}, d_E)$, et préciser son rapport de contraction k.

b) Déterminer le point fixe $\overline{x}$ de f directement, en résolvant $f(\overline{x}) = \overline{x}$.

c) En partant de $x_0 = 0$, calculer x_1, puis utiliser la proposition 2.7 pour majorer $d(x_3, \overline{x})$.

Exercice 2.6. Utiliser la méthode de Newton pour approcher $\sqrt{3}$ en l'appliquant à $f(x) = x^2 - 3$.

a) Montrer que l'itération de Newton associée à f s'écrit $g(x) = \dfrac{1}{2}\left(x + \dfrac{3}{x}\right)$.

b) En partant de $x_0 = 2$, calculer x_1 et x_2, et comparer le nombre de décimales exactes obtenues à chaque étape avec $\sqrt{3} \approx 1{,}732\,050\,807\,6$.

Exercice 2.7. Les notions d'espace métrique, de complétude et de point fixe ne sont pas enseignées au secondaire. Certaines idées de ce chapitre peuvent toutefois y être exploitées.

a) La formule de la distance entre deux points est étudiée en quatrième secondaire. Proposer une activité, fondée sur la distance de Manhattan (exemple 2.2), qui amène les élèves à questionner cette formule. Quel est l'objectif d'apprentissage visé ?

b) Expliquer comment l'itération $x_{n+1} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right)$ de l'exercice 2.6 pourrait être utilisée au premier cycle du secondaire, lors de l'étude de la racine carrée. Quelle notion de ce chapitre l'enseignant mobilise-t-il sans la nommer ?

c) Relever un risque de confusion pour les élèves si l'on emploie le mot « distance » dans deux sens différents.

3 Fractales et systèmes itératifs de fonctions

Les notions des deux chapitres précédents vont maintenant être utilisées pour étudier les fractales. Au chapitre 1, nous avons décrit les transformations affines du plan à l'aide de six nombres et étudié les similitudes et les contractions. Au chapitre 2, nous avons muni l'ensemble des compacts non vides du plan d'une distance, obtenant ainsi l'espace $(\mathcal{K}(\mathbb{R}^2), d_H)$, puis démontré le théorème du point fixe de Banach. Nous verrons comment quelques transformations affines permettent de définir des figures comme un flocon au contour infiniment dentelé ou une fougère. L'existence de ces figures sera justifiée par le théorème de Banach, appliqué à l'espace des compacts du plan.

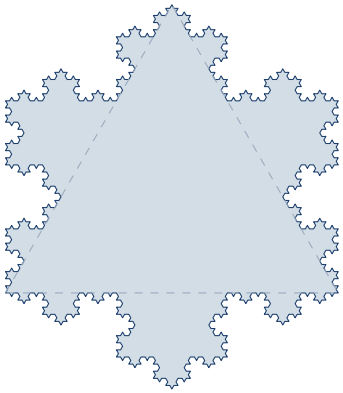


FIGURE 3.1 : Le *flocon de von Koch*, à la quatrième étape de sa construction, superposé au triangle équilatéral dont il est issu (en pointillé). À la limite, son contour a une longueur infinie, alors qu'il délimite une région d'aire finie : cette propriété en fait un exemple classique de *fractale*.

Benoît Mandelbrot a introduit le mot *fractale* en 1975, à partir du latin *fractus*, qui signifie « brisé » [Man75]. Les objets ainsi désignés présentent des irrégularités à toutes les échelles : l'agrandissement ne les fait pas apparaître lisses. Plusieurs étaient pourtant connus depuis la fin du XIXe siècle ou le début du XXe siècle. Ils avaient été construits pour fournir des *contre-exemples*, comme des courbes continues sans tangente ou des ensembles comportant une infinité de trous. Leurs propriétés leur avaient valu le nom de *monstres*. Avec le développement de l'informatique, il est devenu plus facile de les représenter et d'étudier leur intérêt pour décrire des formes naturelles : côtes, nuages, végétaux ou réseaux sanguins, que les figures de la géométrie classique représentent moins bien.

Nous étudierons d'abord la courbe de von Koch et le triangle de Sierpiński. Ces exemples permettront d'introduire l'*autosimilarité* et la *dimension de similitude*. Nous définirons ensuite les *systèmes itératifs de fonctions* (SIF) et montrerons comment le théorème de Banach justifie les constructions qui leur sont associées. Après avoir présenté deux algorithmes de dessin, nous reviendrons sur les travaux de Karl Weierstrass et de Benoît Mandelbrot pour situer l'évolution de ces objets, depuis les contre-exemples de l'analyse

jusqu'aux modèles de phénomènes naturels et financiers.

3.1 Des monstres mathématiques

3.1.1 La courbe de von Koch

Considérons d'abord un segment de longueur 1, que nous appellerons l'étape K_0. Pour obtenir l'étape K_1, on partage ce segment en trois parties égales, on remplace la partie du milieu par les deux autres côtés d'un triangle équilatéral construit sur elle, et on obtient une ligne brisée formée de quatre segments de longueur $\frac{1}{3}$. On recommence ensuite la même opération sur chacun des quatre segments, ce qui donne K_2, formée de 16 segments de longueur $\frac{1}{9}$, et ainsi de suite (figure 3.2). La **courbe de von Koch** K est la figure limite de ce processus, lorsque le nombre d'étapes tend vers l'infini. Nous verrons à la section 3.4 comment justifier l'existence de cette limite dans l'espace $(\mathcal{K}(\mathbb{R}^2), d_H)$.

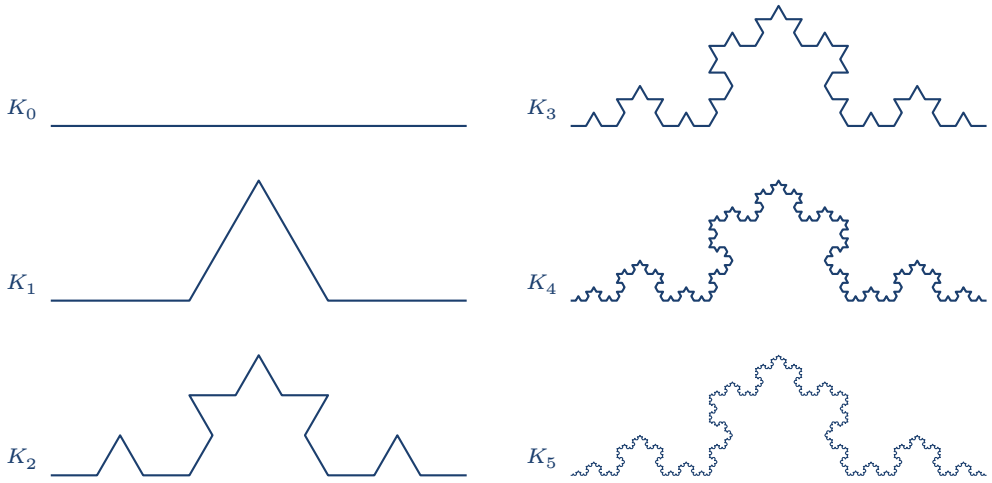


FIGURE 3.2 : Les premières étapes de la construction de la courbe de von Koch. À chaque étape, chaque segment est remplacé par quatre segments trois fois plus courts ; dès l'étape K_5, formée de $4^5 = 1024$ segments, la ligne brisée se distingue difficilement de la courbe limite à l'échelle du dessin.

Deux propriétés de la courbe de von Koch la distinguent des courbes habituellement étudiées au secondaire et en calcul différentiel.

Sa longueur est infinie. À chaque étape, chaque segment est remplacé par quatre segments dont la longueur est le tiers de la sienne : la longueur totale est donc multipliée par $\frac{4}{3}$. La longueur de K_n est ainsi

$$L_n = \left(\frac{4}{3}\right)^n,$$

le terme général d'une suite géométrique de raison $\frac{4}{3} > 1$, qui tend vers l'infini. Aucune longueur finie ne peut donc être attribuée à K : la courbe limite est plus longue que chacune des lignes brisées K_n, dont les longueurs dépassent toute borne. Pour donner une idée de cette croissance, on a $L_{10} \approx 17{,}8$, $L_{50} \approx 1{,}8 \times 10^6$. Si le segment de départ mesurait un centimètre, K_{50} serait plus longue que la distance de Québec à Montréal, bien que la figure puisse être dessinée sur un timbre-poste.

Elle n'a de tangente en aucun point. Une courbe lisse, comme le graphe d'une fonction dérivable, finit toujours par ressembler à un segment de droite, sa tangente, quand on l'observe de suffisamment près. Pour la courbe de von Koch, l'agrandissement reproduit au contraire les mêmes irrégularités : le petit morceau encadré à la figure 3.3, agrandi trois fois, est une copie exacte de la courbe entière, avec ses pointes et ses creux. On peut démontrer qu'en aucun point de K il n'existe de tangente ; c'est la propriété que von Koch cherchait à illustrer (note historique 3.1).

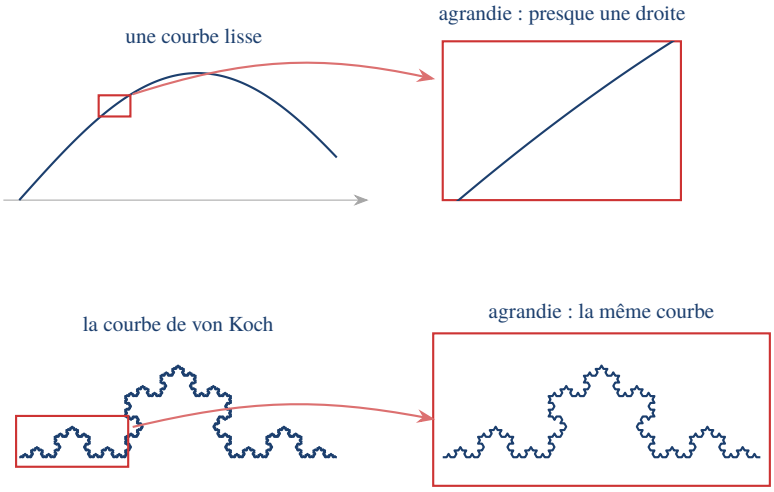


FIGURE 3.3 : L'agrandissement d'une courbe lisse (en haut) fait apparaître localement sa tangente. Agrandir trois fois le premier quart de la courbe de von Koch (en bas) reproduit exactement la courbe de départ : aucune échelle d'observation ne la rend lisse.

En disposant trois courbes de von Koch sur les côtés d'un triangle équilatéral, pointes vers l'extérieur, on obtient le **flocon de von Koch** de la figure 3.1. Son contour, formé de trois courbes de longueur infinie, est lui aussi de longueur infinie. L'aire qu'il délimite, au contraire, est finie : le flocon tient dans n'importe quel disque qui contient le triangle de départ. Le calcul exact de cette aire se fait à l'aide d'une série géométrique (exercice 3.1).

Note historique 3.1 Helge von Koch



Helge von Koch (1870–1924). Domaine public, *Wikimedia Commons*.

Niels Fabian Helge von Koch (1870–1924) est né à Stockholm, dans une famille de la noblesse suédoise. Élève de Gösta Mittag-Leffler, lui-même disciple de Karl Weierstrass, il soutient en 1892 une thèse sur les déterminants infinis et les équations différentielles linéaires, sujets de ses premiers travaux. Il enseigne ensuite à l'Institut royal de technologie de Stockholm, puis succède en 1911 à Mittag-Leffler à l'université de Stockholm. En théorie des nombres, il établit en 1901 le résultat suivant : l'hypothèse de Riemann, le plus célèbre problème ouvert des mathématiques, équivaut à une estimation précise de l'erreur commise quand on compte les nombres premiers à l'aide du logarithme intégral.

La courbe qui porte son nom paraît en 1904, dans un article intitulé : « Sur une courbe continue sans tangente, obtenue par une construction géométrique élémentaire » [Koc04]. Trente ans plus tôt, Weierstrass avait exhibé une fonction continue qui n'est dérivable nulle part (section 3.7.1), mais son exemple, défini par une série

de fonctions trigonométriques, se prêtait moins facilement à une interprétation géométrique. Von Koch souhaitait présenter ce phénomène à l'aide d'une construction géométrique que l'on puisse effectuer à la règle et au compas. Trois quarts de siècle plus tard, son flocon allait devenir l'une des figures les plus connues de la géométrie fractale [ORe].

3.1.2 Le triangle et le tapis de Sierpiński

Le deuxième exemple repose sur le retrait successif de parties d'une figure. Partons d'un triangle équilatéral plein S_0, de côté 1. Relions les milieux de ses trois côtés : on le découpe ainsi en quatre triangles équilatéraux de côté $\frac{1}{2}$, et on retire (l'intérieur de) celui du centre, qui pointe vers le bas. Il reste trois triangles pleins, qui forment S_1. On répète l'opération dans chacun d'eux pour obtenir S_2, formé de 9 triangles de côté $\frac{1}{4}$, et ainsi de suite (figure 3.4). Le **triangle de Sierpiński** S est la limite de la suite $S_0, S_1, S_2, \ldots$:

$$S = \lim_{n\to\infty} S_n.$$

Il faut toutefois préciser le sens de cette limite, puisque les S_n sont des ensembles et non des nombres. C'est précisément ce que permet la distance de Hausdorff du chapitre 2 : les S_n sont des compacts du plan, donc des « points » de l'espace métrique $(\mathcal{K}(\mathbb{R}^2), d_H)$, et dire que S_n tend vers S, c'est dire que $d_H(S_n, S) \to 0$, au sens de la définition 2.10. Nous démontrerons à la section 3.4 que cette limite existe. Comme chaque étape est contenue dans la précédente ($S_{n+1} \subseteq S_n$), on peut aussi décrire la limite de la façon suivante : c'est l'ensemble des points qui ne sont jamais retirés, c'est-à-dire l'intersection $S_0 \cap S_1 \cap S_2 \cap \cdots$ de toutes les étapes.



FIGURE 3.4 : Construction du triangle de Sierpiński par retraits successifs. L'étape S_n est formée de 3^n triangles pleins de côté $\left(\frac{1}{2}\right)^n$.

À chaque étape, chaque triangle plein est remplacé par trois triangles dont les côtés sont deux fois plus courts, donc dont l'aire est quatre fois plus petite (selon le facteur k^2 de la proposition 1.11). L'aire est donc multipliée par $\frac{3}{4}$ à chaque étape :

$$\text{Aire}(S_n) = \left(\frac{3}{4}\right)^n \text{Aire}(S_0) = \left(\frac{3}{4}\right)^n \frac{\sqrt{3}}{4} \xrightarrow[n\to\infty]{} 0.$$

Puisque S est contenu dans chacun des S_n, son aire est plus petite que toutes ces quantités : elle est nulle. Le triangle de Sierpiński n'est pas vide pour autant. Il contient les trois sommets du triangle de départ, les sommets de tous les petits triangles de toutes les étapes, et les côtés de tous ces triangles, qu'aucune étape ne vient jamais retirer. Il contient donc une infinité de segments, dont la longueur totale est infinie : les 3^n triangles de S_n ont des côtés de longueur $\left(\frac{1}{2}\right)^n$, et leurs contours, qui sont tous contenus dans S, totalisent déjà une longueur de $3^n \cdot 3 \cdot \left(\frac{1}{2}\right)^n = 3\left(\frac{3}{2}\right)^n$, une quantité qui tend vers l'infini. Ainsi, ni l'aire ni la longueur ne donnent une mesure finie et non nulle du triangle de Sierpiński. La dimension de similitude, introduite à la section 3.3, permettra de mieux comprendre cette situation.

En appliquant un procédé analogue à un carré, on obtient le **tapis de Sierpiński** : on partage le carré unité en neuf carrés de côté $\frac{1}{3}$, on retire celui du centre, et on recommence dans chacun des huit carrés restants (figure 3.5). À chaque étape, l'aire est multipliée par $\frac{8}{9}$; elle tend donc vers zéro, elle aussi.

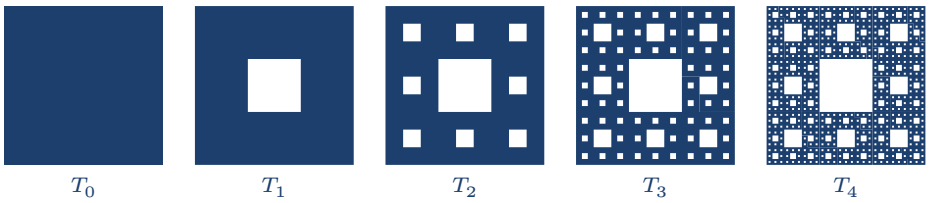
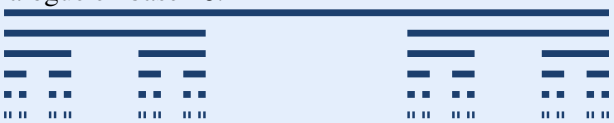


FIGURE 3.5 : Construction du tapis de Sierpiński par retraits successifs. L'étape T_n est formée de 8^n carrés pleins de côté $\left(\frac{1}{3}\right)^n$; son aire, $\left(\frac{8}{9}\right)^n$, tend vers zéro.

Remarque 3.1 **L'ensemble de Cantor**

Une construction analogue sur un segment avait été proposée antérieurement. Retirons du segment $[0, 1]$ son tiers central ouvert $]\frac{1}{3}, \frac{2}{3}[$, puis le tiers central de chacun des deux segments restants, et ainsi de suite. L'ensemble obtenu à la limite est l'**ensemble de Cantor**, décrit par Georg Cantor en 1883. Sa longueur totale est nulle, puisqu'à l'étape n il reste 2^n segments de longueur $\left(\frac{1}{3}\right)^n$, soit une longueur $\left(\frac{2}{3}\right)^n \to 0$. Il contient pourtant une infinité de points, et même « autant » de points que le segment entier : on démontre qu'il est formé exactement des nombres de $[0, 1]$ qui s'écrivent en base 3 sans utiliser le chiffre 1. L'exercice 3.3 propose une construction analogue en base 10.



Note historique 3.2 **Wacław Sierpiński**



Wacław Sierpiński (1882–1969), en 1928. Domaine public, *Wikimedia Commons*.

Wacław Sierpiński (1882–1969) est né à Varsovie, alors sous domination russe. Il y étudie les mathématiques auprès du mathématicien russe Gueorgui Voronoï, enseigne les mathématiques et la physique dans une école secondaire de filles, puis obtient son doctorat à Cracovie en 1906. Nommé à l'université de Lwów, il y donne en 1909 l'un des tout premiers cours universitaires entièrement consacrés à la théorie des ensembles, dont les fondements faisaient encore l'objet de débats. Surpris en Russie par la Première Guerre mondiale, il est interné comme ressortissant ennemi, d'abord à Viatka, puis à Moscou, où il poursuit ses recherches avec Nikolaï Louzine.

C'est durant cette période qu'il publie le triangle (1915) et le tapis (1916) qui portent son nom [Sie15 ; Sie16]. Ces constructions lui permettaient d'étudier certaines propriétés topologiques : le triangle est une courbe dont chaque point est un point de « ramification », d'où partent plusieurs branches, et le tapis est *universel*, en ce sens qu'il contient une copie de toute courbe plane imaginable. De retour à Varsovie en 1918, Sierpiński devient l'un des principaux représentants de l'école mathématique

polonaise de l'entre-deux-guerres. Avec Stefan Mazurkiewicz, et sur une idée de Zygmunt Janiszewski, il fonde en 1920 la revue *Fundamenta Mathematicae*, consacrée à la théorie des ensembles et à la topologie, où Stefan Banach (note historique 2.3) publiera plusieurs de ses travaux. Pendant l'occupation allemande, il enseigne dans l'université clandestine de Varsovie, et perd sa bibliothèque et ses manuscrits dans l'incendie de sa maison, en 1944. Auteur de plus de 700 articles et d'une cinquantaine de livres, il a aussi laissé son nom à une famille de nombres entiers, les *nombres de Sierpiński* [ORj].

Pour la classe 3.1 **Les suites géométriques dans les constructions fractales**

Les calculs précédents donnent un sens géométrique à la raison d'une suite : la longueur est multipliée par $\frac{4}{3}$, ou l'aire par $\frac{3}{4}$. En quatrième ou cinquième secondaire, on peut demander aux élèves de construire les étapes K_0 à K_3 de la courbe de von Koch sur du papier pointé isométrique. Ils peuvent ensuite calculer les longueurs, reconnaître une suite géométrique et prévoir son comportement. Le fait que la longueur augmente sans borne alors que la figure reste dans une région bornée fournit une occasion de discuter de la notion de limite avant d'en donner une définition formelle.

Outil 3.1 **Construire la courbe de von Koch dans GeoGebra**

GeoGebra permet de créer un *outil personnalisé* à partir d'une construction : on part d'un segment $[AB]$, on place les points situés au tiers et aux deux tiers, on construit le sommet du triangle équilatéral par une rotation de 60°, puis on trace les quatre segments de K_1. Le menu *Outils* → *Créer un nouvel outil* transforme cette construction en un outil qui prend deux points et produit les quatre segments. Appliqué à chacun des segments obtenus, il donne K_2, puis K_3. L'activité rend concrète l'idée de *remplacer chaque segment par une copie réduite de la figure*, qui est au cœur des SIF (section 3.4), et elle montre rapidement pourquoi on a besoin d'un programme pour aller plus loin : le nombre de segments est multiplié par 4 à chaque étape.

3.2 Autosimilarité

Revenons à la figure 3.3 : le premier quart de la courbe de von Koch, agrandi trois fois, redonne la courbe entière. Il en va de même des trois autres quarts, à condition de les faire aussi tourner. Autrement dit, la courbe K est formée de *quatre copies réduites d'elle-même*. De même, le triangle de Sierpiński est formé de trois copies de lui-même, deux fois plus petites, placées dans ses trois coins. Ces observations concernent les figures obtenues à la limite. Pour les exprimer plus précisément, utilisons le langage du chapitre 1. Rappelons qu'une similitude de rapport r multiplie toutes les distances par r (définition 1.11) ; lorsque $0 < r < 1$, c'est en particulier une contraction de rapport r au sens de la définition 2.14.

Définition 3.1 **Ensemble autosimilaire**

Soit E un sous-ensemble compact non vide du plan. On dit que E est **autosimilaire** s'il existe un nombre fini $N \geq 2$ de similitudes $w_1, w_2, \dots, w_N$, de rapports $r_1, r_2, \dots, r_N$ compris strictement entre 0 et 1, telles que

$$E = w_1(E) \cup w_2(E) \cup \cdots \cup w_N(E).$$

Chacun des ensembles $w_i(E)$ est une **copie réduite** de E.

La définition exige que E soit *exactement* la réunion de ses copies : les deux ensembles doivent être égaux. Elle n'interdit pas, en revanche, que deux copies se touchent, ni même qu'elles se chevauchent ; nous reviendrons sur ce point à la section 3.5. Examinons d'abord cette propriété sur des figures familières.

Exemple 3.1 **Le segment et le carré**

Le segment $[0,1]$ est la réunion de ses deux moitiés, $h_{1/2}([0,1]) = [0, \frac{1}{2}]$ et $h_{1/2}([0,1]) + \frac{1}{2} = [\frac{1}{2}, 1]$: il est autosimilaire, avec deux similitudes de rapport $\frac{1}{2}$ (il l'est aussi avec trois similitudes de rapport $\frac{1}{3}$, ou m de rapport $\frac{1}{m}$). Pour le carré unité, on a de même que $C = [0,1] \times [0,1]$ est la réunion de quatre carrés de côté $\frac{1}{2}$: $C = w_1(C) \cup w_2(C) \cup w_3(C) \cup w_4(C)$, où

$$w_i \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \mathbf{t}_i,$$

les vecteurs $\mathbf{t}_i$ étant $(0,0)$, $(\frac{1}{2}, 0)$, $(0, \frac{1}{2})$ et $(\frac{1}{2}, \frac{1}{2})$.

Exemple 3.2 **Tout triangle est autosimilaire**

Soit T un triangle plein quelconque, de sommets A, B et C. En joignant les milieux de ses côtés, on le découpe en quatre triangles semblables à T, de rapport $\frac{1}{2}$ (figure 3.6). Les trois triangles situés aux sommets sont les images de T par les homothéties $h_{A,1/2}$, $h_{B,1/2}$ et $h_{C,1/2}$ de la proposition 1.5. Le quatrième, celui du centre, a subi un demi-tour : c'est l'image de T par l'homothétie $h_{G,-1/2}$ de centre le centre de gravité G de T et de rapport $-\frac{1}{2}$, qui combine une réduction de moitié et un demi-tour. C'est donc une similitude de rapport $\frac{1}{2}$, et $T = h_{A,1/2}(T) \cup h_{B,1/2}(T) \cup h_{C,1/2}(T) \cup h_{G,-1/2}(T)$.

Ces exemples montrent que l'autosimilarité ne suffit pas à caractériser une fractale : les carrés et les triangles possèdent aussi cette propriété. Examinons maintenant les deux constructions introduites au début du chapitre.

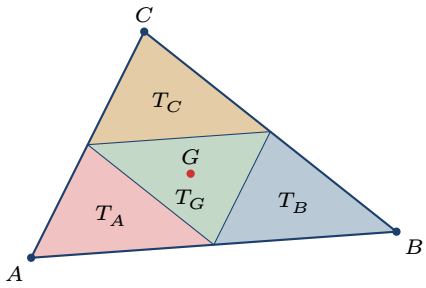


FIGURE 3.6 : Décomposition d'un triangle T en quatre copies de rapport un demi : les trois copies de coin $T_A = h_{A,1/2}(T)$, $T_B = h_{B,1/2}(T)$ et $T_C = h_{C,1/2}(T)$, images de T par des homothéties de rapport $\frac{1}{2}$ centrées aux sommets, et la copie centrale $T_G = h_{G,-1/2}(T)$, image de T par l'homothétie de rapport $-\frac{1}{2}$ centrée au centre de gravité G.

Exemple 3.3 **Le triangle de Sierpiński**

Notons $S_1 = (0,0)$, $S_2 = (1,0)$ et $S_3 = (\frac{1}{2}, \frac{\sqrt{3}}{2})$ les sommets du triangle de départ, et $w_i = h_{S_i,1/2}$ l'homothétie de centre S_i et de rapport $\frac{1}{2}$. Par la proposition 1.5, $w_i(P) = \frac{1}{2}P + \frac{1}{2}S_i$, c'est-à-dire

$$w_1\begin{pmatrix}x\\y\end{pmatrix} = \frac{1}{2}\begin{pmatrix}x\\y\end{pmatrix}, \qquad w_2\begin{pmatrix}x\\y\end{pmatrix} = \frac{1}{2}\begin{pmatrix}x\\y\end{pmatrix} + \begin{pmatrix}\frac{1}{2}\\0\end{pmatrix},$$
$$w_3\begin{pmatrix}x\\y\end{pmatrix} = \frac{1}{2}\begin{pmatrix}x\\y\end{pmatrix} + \begin{pmatrix}\frac{1}{4}\\\frac{\sqrt{3}}{4}\end{pmatrix}.$$

Le triangle de Sierpiński vérifie $S = w_1(S) \cup w_2(S) \cup w_3(S)$ (figure 3.7, à gauche) : il est autosimilaire, avec trois similitudes de rapport $\frac{1}{2}$. Il l'est aussi avec neuf similitudes de rapport $\frac{1}{4}$, les composées $w_i \circ w_j$, en décomposant à son tour chacune des trois copies en trois parties semblables.

Exemple 3.4 **La courbe de von Koch**

La courbe de von Koch, construite sur le segment joignant $(0,0)$ à $(1,0)$, est la réunion de quatre copies d'elle-même, de rapport $\frac{1}{3}$ (figure 3.7, à droite). Les deux copies extrêmes sont obtenues par une homothétie de rapport $\frac{1}{3}$, suivie ou non d'une translation ; les deux copies du centre, qui forment la « pointe », sont en plus tournées de $\pm 60°$. En utilisant les matrices de rotation r_θ de la définition 1.7, on trouve

$$w_1\begin{pmatrix}x\\y\end{pmatrix} = \frac{1}{3}\begin{pmatrix}x\\y\end{pmatrix}, \qquad w_2\begin{pmatrix}x\\y\end{pmatrix} = \frac{1}{3}r_{\pi/3}\begin{pmatrix}x\\y\end{pmatrix} + \begin{pmatrix}\frac{1}{3}\\0\end{pmatrix},$$
$$w_3\begin{pmatrix}x\\y\end{pmatrix} = \frac{1}{3}r_{-\pi/3}\begin{pmatrix}x\\y\end{pmatrix} + \begin{pmatrix}\frac{1}{2}\\\frac{\sqrt{3}}{6}\end{pmatrix}, \qquad w_4\begin{pmatrix}x\\y\end{pmatrix} = \frac{1}{3}\begin{pmatrix}x\\y\end{pmatrix} + \begin{pmatrix}\frac{2}{3}\\0\end{pmatrix}.$$

On vérifie par exemple que w_2 envoie les extrémités $(0,0)$ et $(1,0)$ de la courbe sur $(\frac{1}{3},0)$ et sur $\frac{1}{3}(\cos\frac{\pi}{3}, \sin\frac{\pi}{3}) + (\frac{1}{3},0) = (\frac{1}{2}, \frac{\sqrt{3}}{6})$, le sommet de la pointe : la copie $w_2(K)$ correspond donc à la partie montante située à gauche. On procède de la même façon pour w_3, qui envoie $(0,0)$ sur la pointe et $(1,0)$ sur $(\frac{2}{3},0)$.

On retrouve une décomposition analogue pour le tapis de Sierpiński. Les huit homothéties de rapport $\frac{1}{3}$ qui envoient le carré unité sur les carrés entourant le carré central donnent

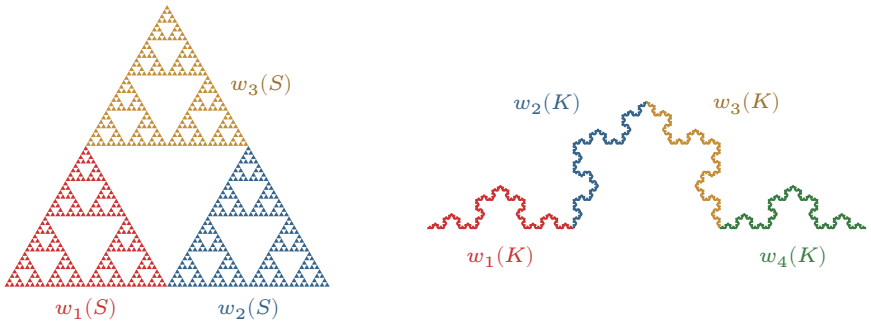


FIGURE 3.7 : À gauche, le triangle de Sierpiński, réunion de ses trois copies $w_1(S)$, $w_2(S)$ et $w_3(S)$; à droite, la courbe de von Koch, réunion de ses quatre copies $w_1(K)$, ..., $w_4(K)$. Les copies colorées sont les images de la figure entière par une similitude de rapport $\frac{1}{2}$ (à gauche) ou $\frac{1}{3}$ (à droite).

huit copies dont la réunion est le tapis. Pour l'ensemble de Cantor, les deux copies de rapport $\frac{1}{3}$ sont obtenues par $\frac{1}{3}x$ et $\frac{1}{3}x + \frac{2}{3}$.

Remarque 3.2 **Une autosimilarité parfois seulement statistique**

Les objets naturels auxquels on compare souvent les fractales (une côte rocheuse, un chou-fleur, un réseau de rivières) ne sont jamais exactement autosimilaires au sens de la définition 3.1. Un morceau de côte agrandi ne reproduit pas exactement la côte entière ; il peut cependant présenter des caractéristiques statistiques comparables. De plus, cette ressemblance ne vaut que sur une plage d'échelles limitée : à l'échelle du grain de sable, la description de la côte par le même modèle n'est plus appropriée. On parle alors d'autosimilarité *statistique*, une notion que Mandelbrot a placée au cœur de ses travaux (section 3.7.2).

Pour la classe 3.2 **Une figure semblable à une partie d'elle-même**

Au secondaire, deux figures sont *semblables* lorsqu'on peut passer de l'une à l'autre par une similitude, et le rapport de similitude intervient dans de nombreux problèmes de mesure. Une figure autosimilaire est simplement une figure semblable à chacune de ses parties. L'exemple 3.2 en fournit une version accessible aux élèves qui étudient les figures semblables. Les élèves savent repérer le milieu des côtés d'un triangle, qu'ils utilisent déjà pour tracer les médianes. En reliant deux à deux ces trois milieux, ils découpent le triangle en quatre petits triangles ; en mesurant les côtés, ils constatent que chacun est semblable au grand, dans un rapport de $\frac{1}{2}$, et que les quatre sont isométriques. Le triangle est donc formé de quatre copies réduites de lui-même. Demander aux élèves de refaire ce découpage dans chacun des petits triangles, puis de noircir celui du centre, permet d'amorcer la construction du triangle de Sierpiński.

3.3 La dimension de similitude

À la section 3.1.2, nous avons constaté que le triangle de Sierpiński a une aire nulle et contient des segments de longueur totale infinie. La courbe de von Koch présente une situation analogue : sa longueur est infinie, mais son aire est nulle. En effet, la courbe

construite sur un segment de longueur ℓ est contenue dans le triangle isocèle qui a ce segment pour base et des angles de 30° à la base, d'aire $\frac{\sqrt{3}}{12}\ell^2$. En appliquant cette remarque à chacun des 4^n segments de K_n, on voit que K est contenue dans une réunion de triangles d'aire totale $4^n \cdot \frac{\sqrt{3}}{12}\left(\frac{1}{3}\right)^{2n} = \frac{\sqrt{3}}{12}\left(\frac{4}{9}\right)^n$, qui tend vers 0. La notion de dimension que nous allons introduire permet de situer ces objets entre les courbes et les surfaces, du point de vue de la mesure.

3.3.1 Compter les copies

Découpons un segment, un carré et un cube en morceaux semblables à l'objet entier, en divisant chaque arête en s parties égales (figure 3.8). On obtient respectivement $N = s$, $N = s^2$ et $N = s^3$ copies, chacune de rapport $r = \frac{1}{s}$. L'exposant qui apparaît est, chaque fois, la dimension usuelle de l'objet :

$$N = s^d, \qquad \text{c'est-à-dire} \qquad N = \left(\frac{1}{r}\right)^d \qquad \text{ou encore} \qquad d = \frac{\ln N}{\ln(1/r)}.$$

Cette relation correspond au facteur k^2 de la remarque 1.6, utilisé dans le sens inverse. Réduire un carré d'un facteur r divise son aire par $(1/r)^2$, et il faut donc $(1/r)^2$ copies réduites pour reconstituer l'aire de départ ; de même, il en faut $(1/r)^3$ pour un cube, puisque les volumes sont multipliés par r^3.

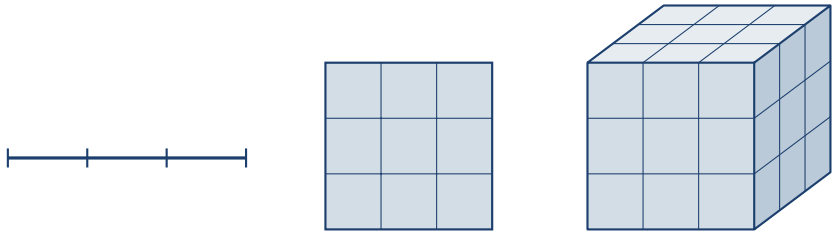


FIGURE 3.8 : En divisant chaque arête en $s = 3$ parties égales, on découpe un segment en 3 copies, un carré en 9 copies et un cube en 27 copies, toutes de rapport $\frac{1}{3}$. Le nombre de copies est s^d, où d est la dimension usuelle de l'objet.

La formule $d = \ln N / \ln(1/r)$ ne fait intervenir que le nombre de copies et leur rapport. Nous pouvons donc l'appliquer à la courbe de von Koch ou au triangle de Sierpiński.

Définition 3.2 **Dimension de similitude**

Soit E un ensemble autosimilaire, réunion de N copies $w_1(E), \ldots, w_N(E)$ qui ont toutes le même rapport $r \in]0, 1[$, et qui ne se chevauchent pas (elles peuvent se toucher, en un sens que la section 3.5 précisera). On appelle **dimension de similitude** de E le nombre

$$\dim_{\text{sim}}(E) = \frac{\ln N}{\ln(1/r)},$$

c'est-à-dire l'unique nombre réel d tel que $Nr^d = 1$.

Exemple 3.5 **Quelques dimensions de similitude**

Ensemble	N	r	$\dim_{\text{sim}}(E) = \ln N / \ln(1/r)$
Ensemble de Cantor	2	$\frac{1}{3}$	$\ln 2/\ln 3 \approx 0{,}6309$
Segment	3	$\frac{1}{3}$	1
Courbe de von Koch	4	$\frac{1}{3}$	$\ln 4/\ln 3 \approx 1{,}2619$
Triangle de Sierpiński	3	$\frac{1}{2}$	$\ln 3/\ln 2 \approx 1{,}5850$
Tapis de Sierpiński	8	$\frac{1}{3}$	$\ln 8/\ln 3 \approx 1{,}8928$
Carré	9	$\frac{1}{3}$	2

Ces valeurs permettent de comparer les ensembles avec des figures familières. La dimension de l'ensemble de Cantor est comprise entre 0, dimension d'un point, et 1, dimension d'un segment. Celle de la courbe de von Koch dépasse la dimension usuelle d'une courbe, tandis que celle du tapis de Sierpiński se rapproche de la dimension d'une surface, malgré son aire nulle.

Une question se pose au sujet de la définition 3.2 : un même ensemble peut se décomposer de plusieurs façons. Le triangle de Sierpiński est formé de 3 copies de rapport $\frac{1}{2}$, mais aussi de 9 copies de rapport $\frac{1}{4}$ (exemple 3.3). Pour ces deux découpages, le calcul donne la même dimension :

$$\frac{\ln 9}{\ln 4} = \frac{\ln 3^2}{\ln 2^2} = \frac{2\ln 3}{2\ln 2} = \frac{\ln 3}{\ln 2}.$$

Plus généralement, si E est formé de N copies de rapport r, en redécoupant chaque copie on obtient N^m copies de rapport r^m, et $\ln N^m / \ln(1/r^m) = \ln N / \ln(1/r)$: la dimension de similitude est la même pour tous ces découpages.

3.3.2 Ce que mesure la dimension

Pour comprendre pourquoi ce nombre est appelé une dimension, nous allons proposer un raisonnement heuristique. Il permet aussi d'interpréter la longueur infinie et l'aire nulle des exemples précédents. Imaginons que nous disposions, pour chaque nombre $d \geq 0$, d'une façon de mesurer la « taille d-dimensionnelle » $M_d(E)$ d'un ensemble E : la longueur pour $d = 1$, l'aire pour $d = 2$, le volume pour $d = 3$. Ces trois mesures ont en commun la façon dont elles réagissent à une similitude de rapport r : elles sont multipliées par r, r^2 et r^3 respectivement. Exigeons la même chose de M_d :

$$M_d(w(E)) = r^d\, M_d(E) \qquad \text{pour toute similitude } w \text{ de rapport } r.$$

Si E est la réunion de N copies de rapport r qui ne se chevauchent pas, la taille de E est la somme des tailles des copies :

$$M_d(E) = N\, r^d\, M_d(E).$$

Examinons les possibilités compatibles avec cette égalité. Si $Nr^d = 1$, c'est-à-dire si $d = \dim_{\text{sim}}(E)$, elle est satisfaite quelle que soit la valeur de $M_d(E)$. Sinon, il faut que $M_d(E)$ soit nulle, ou infinie. Pour la courbe de von Koch, par exemple :

- avec $d = 1$ (la longueur), $Nr^d = 4 \cdot \frac{1}{3} = \frac{4}{3} \neq 1$; la longueur doit donc être nulle ou infinie, et elle ne peut être nulle, puisque K contient deux points à distance 1. Elle est infinie, comme nous l'avons calculé ;
- avec $d = 2$ (l'aire), $Nr^d = 4 \cdot \frac{1}{9} = \frac{4}{9} \neq 1$; l'aire doit donc être nulle ou infinie, et elle ne peut être infinie, puisque K est borné. Elle est nulle ;
- la seule « mesure » susceptible de donner une valeur finie et non nulle à K est celle de dimension $d = \ln 4/\ln 3$.

On peut faire une analogie avec les unités habituelles : chercher à mesurer une surface au moyen d'une longueur conduit à une valeur infinie, tandis que l'aire d'un fil idéal est nulle. La dimension de similitude est l'unique exposant pour lequel la mesure n'est ni nulle ni infinie.

Remarque 3.3 **La dimension de Hausdorff**

Le raisonnement précédent demeure heuristique : nous n'avons pas construit les mesures M_d. Felix Hausdorff (note historique 2.2) les a construites en 1918, pour tous les sous-ensembles de $\mathbb{R}^n$ et tous les $d \geq 0$, sans supposer que les ensembles sont autosimilaires. Il en tire la **dimension de Hausdorff**, qui a un sens pour n'importe quel ensemble. Pour les ensembles autosimilaires dont les copies ne se chevauchent pas, elle coïncide avec la dimension de similitude [Hut81] : c'est ce qui justifie, a posteriori, le raisonnement ci-dessus. La définition de Hausdorff dépasse le cadre de ce cours ; on la trouvera, par exemple, dans [Fal14].

Nous pouvons maintenant préciser ce que nous entendrons par « fractale ». Mandelbrot a proposé de réserver ce nom aux ensembles dont la dimension de Hausdorff dépasse strictement la dimension topologique (qui vaut 0 pour un ensemble de points isolés, 1 pour une courbe, 2 pour une surface) [Man82]. Il a lui-même reconnu que cette définition excluait des objets qu'on voudrait appeler fractals, et aucune définition entièrement satisfaisante n'a été retenue depuis. Pour nos besoins, une **fractale** sera un ensemble irrégulier à toutes les échelles, dont la forme se reproduit, exactement ou approximativement, dans ses parties, et dont la dimension est typiquement fractionnaire. Il faut cependant conserver cette réserve : le tétraèdre de Sierpiński est une fractale de dimension exactement 2 (exercice 3.5).

Pour la classe 3.3 **Logarithmes et échelles de mesure**

Le calcul d'une dimension de similitude est une application concrète de la résolution d'équations exponentielles enseignée en cinquième secondaire. L'équation $3^d = 4$ n'a pas de solution entière ; on la résout en passant aux logarithmes, dans la base de son choix, on peut ainsi vérifier que $\log 4 / \log 3 = \ln 4 / \ln 3 = \log_3 4$, ce qui illustre la formule de changement de base. La dimension de similitude est un *exposant* : elle répond à la question « à quelle puissance faut-il élever le facteur de réduction pour obtenir le nombre de copies ? ».
L'utilisation d'un exposant comme mesure se retrouve dans les *échelles logarithmiques*, que les élèves rencontrent en sciences sans toujours en saisir le fonctionnement. La magnitude d'un séisme sur l'échelle de Richter est le logarithme en base 10 de l'amplitude des ondes enregistrées par un sismographe, comparée à une amplitude de référence : un séisme de magnitude 6 produit des ondes dix fois plus amples qu'un séisme de magnitude 5, et libère environ $10^{1,5} \approx 32$ fois plus d'énergie. Le pH d'une solution et le niveau sonore en décibels fonctionnent de la même façon. Dans chaque cas, le logarithme ramène à une échelle commode des grandeurs qui varient sur plusieurs ordres de grandeur, et il transforme les multiplications en additions : ajouter 1 à la magnitude, c'est multiplier l'amplitude par 10. Faire calculer aux élèves la magnitude de quelques séismes historiques, puis la dimension de quelques fractales, permet de rapprocher deux applications d'un même outil mathématique.

3.4 Systèmes itératifs de fonctions

Jusqu'à présent, nous avons reconnu des copies semblables à l'intérieur d'une figure donnée. Posons maintenant la question dans l'autre sens. Si on se donne des contractions $w_1, \dots, w_N$ du plan, existe-t-il un ensemble E tel que $E = w_1(E) \cup \cdots \cup w_N(E)$? Peut-on établir son unicité et proposer une méthode pour le construire ? Le théorème du point fixe de Banach permet de répondre à ces questions, à condition de l'appliquer à l'ensemble $\mathcal{K}(\mathbb{R}^2)$ des compacts non vides du plan, muni de la distance de Hausdorff introduite au chapitre 2.

Définition 3.3 **Système itératif de fonctions**

Un **système itératif de fonctions** (SIF) du plan est une famille finie $\{w_1, w_2, \dots, w_N\}$ de contractions $w_i : \mathbb{R}^2 \to \mathbb{R}^2$, de rapports respectifs $k_1, k_2, \dots, k_N$. Le nombre $k = \max\{k_1, \dots, k_N\}$ s'appelle le **rapport de contraction** du SIF. On associe au SIF la fonction W qui, à un ensemble compact non vide B, fait correspondre la réunion de ses images :

$$W(B) = w_1(B) \cup w_2(B) \cup \cdots \cup w_N(B).$$

On appelle W l'**opérateur de Hutchinson** du SIF.

Dans ce cours, les w_i seront toujours des applications affines (définition 1.3). L'opérateur de Hutchinson agit sur des *figures*. À partir d'une figure B, il forme N images par les applications du système, puis prend leur réunion. C'est un « photocopieur à plusieurs lentilles », selon l'image de Barnsley [Bar93] : chaque lentille réduit et déplace l'original, et la page de sortie superpose toutes les réductions. En anglais, on parle d'*iterated function system* (IFS).

Remarque 3.4 **Le rapport de contraction d'une application affine**

Pour une similitude, le rapport de contraction est simplement son rapport de similitude. Pour une application affine $w(\mathbf{x}) = A\mathbf{x} + \mathbf{t}$ quelconque, c'est le plus grand facteur par lequel A peut allonger un vecteur, c'est-à-dire le maximum de $\|A\mathbf{u}\|$ lorsque $\mathbf{u}$ parcourt les vecteurs unitaires. Pour une matrice diagonale $\begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}$, c'est $\max\{|a|, |b|\}$: on retrouve le rapport $\frac{2}{3}$ de l'exemple 2.10. Plus généralement, si A est symétrique, c'est la plus grande valeur absolue de ses valeurs propres. Cette caractérisation ne s'étend toutefois pas à une matrice quelconque : les valeurs propres de A ne suffisent pas. Le cisaillement $\begin{pmatrix} \frac{1}{2} & 1 \\ 0 & \frac{1}{2} \end{pmatrix}$ a pour seule valeur propre $\frac{1}{2}$, mais il envoie le vecteur unitaire $(0, 1)$ sur $(1, \frac{1}{2})$, de norme $\frac{\sqrt{5}}{2} > 1$: ce n'est pas une contraction. Dans le cas général, le rapport cherché est la racine carrée de la plus grande valeur propre de la matrice symétrique $A^{\mathsf{T}}A$, mais nous n'en aurons pas besoin.

3.4.1 L'opérateur de Hutchinson est une contraction

Pour appliquer le théorème de Banach à W, nous devons vérifier deux propriétés : que W envoie bien $\mathcal{K}(\mathbb{R}^2)$ dans lui-même, et que c'est une contraction pour la distance de Hausdorff. Pour le premier point, nous utiliserons sans démonstration les propriétés

suivantes : une contraction est continue (voir la démonstration du lemme 2.6), l'image d'un compact par une fonction continue est un compact, et une réunion finie de compacts (donc d'ensembles fermés et bornés) est encore un compact. Le second est l'objet du théorème suivant, dont la démonstration, fondée sur les ε-voisinages de la définition 2.8, est présentée à l'annexe C.

Théorème 3.1 **L'opérateur de Hutchinson est une contraction**

Soit $\{w_1, \dots, w_N\}$ un SIF de rapport de contraction k. Alors son opérateur de Hutchinson W est une contraction de rapport k de l'espace métrique $(\mathcal{K}(\mathbb{R}^2), d_H)$:

$$d_H(W(A), W(B)) \leq k\, d_H(A, B) \qquad \text{pour tous } A, B \in \mathcal{K}(\mathbb{R}^2).$$

Voir l'annexe C.

3.4.2 Le théorème de Hutchinson

Il reste à vérifier la complétude de l'espace pour pouvoir appliquer le théorème de Banach. Nous admettrons le résultat suivant, dû à Hausdorff ; une démonstration se trouve notamment dans [Bar93].

Théorème 3.2 **Complétude de l'espace des compacts**

L'espace métrique $(\mathcal{K}(\mathbb{R}^2), d_H)$ est complet.

Nous pouvons maintenant réunir les résultats précédents pour énoncer le théorème de John Hutchinson [Hut81]. Il donne le lien recherché entre un système de contractions et une figure invariante.

Théorème 3.3 **Hutchinson, 1981**

Soit $\{w_1, \dots, w_N\}$ un SIF du plan, de rapport de contraction k, et soit W son opérateur de Hutchinson. Alors :

(i) il existe un unique compact non vide $A \subseteq \mathbb{R}^2$ tel que

$$A = W(A) = w_1(A) \cup w_2(A) \cup \cdots \cup w_N(A);$$

(ii) pour tout compact non vide B, la suite de compacts $B, W(B), W^2(B), \dots$ converge vers A au sens de la distance de Hausdorff ;

(iii) on dispose de plus de la majoration

$$d_H(W^n(B), A) \leq \frac{k^n}{1-k}\, d_H(B, W(B)).$$

L'ensemble A s'appelle l'**attracteur** du SIF.

Voir l'annexe C.

La démonstration, présentée à l'annexe C, est une application directe du théorème de Banach : l'opérateur W est une contraction (théorème 3.1) de l'espace complet $(\mathcal{K}(\mathbb{R}^2), d_H)$ (théorème 3.2). La démonstration est courte parce qu'elle s'appuie sur les résultats développés dans les deux premiers chapitres. Il faut cependant préciser ce que représente ici le point fixe. Le « point fixe » dont Banach garantit l'existence n'est pas ici un point du

plan, mais une figure entière, l'attracteur ; et la suite qui y converge n'est pas une suite de points, mais une suite de figures. Le point (ii) affirme de plus que la limite ne dépend pas de la figure de départ : c'est ce qu'illustre la figure 3.9, où le SIF du triangle de Sierpiński, appliqué à un carré, au motif « F » du chapitre 1 ou à un disque, fait apparaître chaque fois le même attracteur.

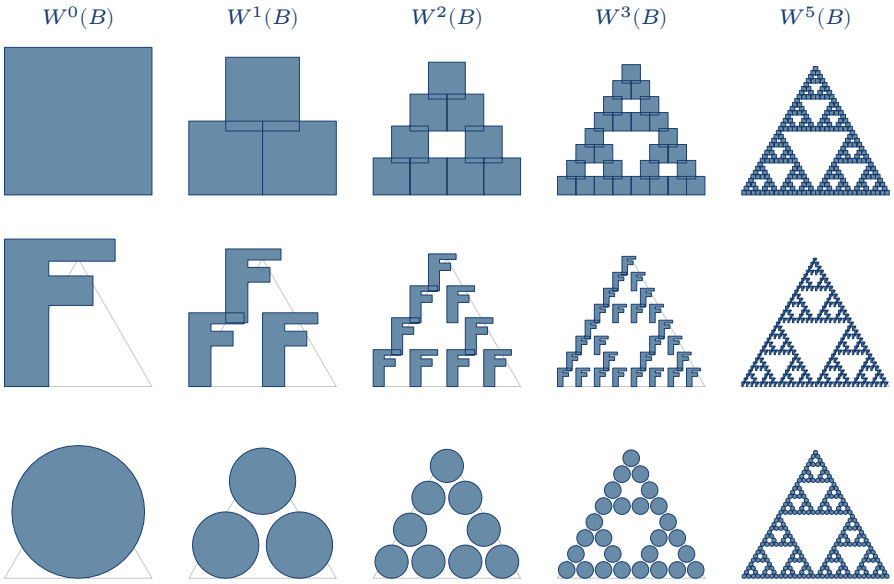


FIGURE 3.9 : Le SIF du triangle de Sierpiński (exemple 3.3) appliqué à trois figures de départ B : un carré, le motif « F » et un disque. Le triangle gris rappelle la position du triangle de départ de la section 3.1.2. Quelle que soit la figure initiale, la suite $W^n(B)$ converge vers le même attracteur, le triangle de Sierpiński : conformément au point (ii) du théorème 3.3.

Remarque 3.5 **Justification des constructions de la section 3.1**

Le théorème 3.3 répond aux deux questions laissées ouvertes à la section 3.1. D'abord, le sens de la « figure limite » : en appliquant l'opérateur de Hutchinson du SIF de l'exemple 3.4 au segment K_0, on obtient exactement la ligne brisée K_1, puis $K_2 = W(K_1)$, et ainsi de suite ; la courbe de von Koch est donc la limite, au sens de d_H, de la suite $K_n = W^n(K_0)$. Il en va de même pour le triangle de Sierpiński, avec $S_n = W^n(S_0)$. Ensuite, l'autosimilarité : un ensemble autosimilaire $E = w_1(E) \cup \cdots \cup w_N(E)$ est un point fixe de l'opérateur W ; par l'unicité du point (i), c'est *l'*attracteur du SIF $\{w_1, \ldots, w_N\}$. Les ensembles autosimilaires sont donc exactement les attracteurs des SIF formés de similitudes.

Exemple 3.6 **Le nombre d'étapes requis pour le dessin**

On veut dessiner le triangle de Sierpiński S sur une page, dans un triangle de 10 cm de côté, à l'aide des triangles pleins de l'étape $S_n = W^n(S_0)$. Déterminons le nombre d'étapes nécessaire pour que l'image ne diffère de S que de moins d'un dixième de millimètre, la taille d'un point d'encre, c'est-à-dire de 0,001 fois le côté. Calculons d'abord $d_H(S_0, W(S_0)) = d_H(S_0, S_1)$. Puisque $S_1 \subseteq S_0$, il suffit de considérer l'éloignement des points de S_0 par rapport à S_1 (formule (2.1)). Le point de S_0 le plus éloigné de S_1 est le centre du triangle central retiré, et sa distance à S_1 est le rayon du cercle inscrit dans ce triangle de côté $\frac{1}{2}$, soit $\frac{1}{2} \cdot \frac{\sqrt{3}}{6} = \frac{\sqrt{3}}{12}$. Avec $k = \frac{1}{2}$, le point (iii) du théorème 3.3 donne

$$d_H(S_n, S) \leq \frac{(1/2)^n}{1 - 1/2} \cdot \frac{\sqrt{3}}{12} = \left(\frac{1}{2}\right)^{n-1} \frac{\sqrt{3}}{12}.$$

Ce majorant vaut environ 0,001 13 pour $n = 8$ et 0,000 56 pour $n = 9$: neuf étapes suffisent, soit $3^9 = 19\,683$ petits triangles à dessiner.

Remarque 3.6 **Les points fixes des w_i appartiennent à l'attracteur**

Chaque w_i possède, par le théorème de Banach, un unique point fixe $\overline{x}_i$, et ce point appartient à l'attracteur A : partant d'un point $a \in A$, la suite $a, w_i(a), w_i^2(a), \ldots$ reste dans A, puisque $w_i(A) \subseteq W(A) = A$, et converge vers $\overline{x}_i$; comme A est fermé, cette limite est dans A. Pour le triangle de Sierpiński, ce sont les trois sommets S_1, S_2 et S_3. En appliquant des composées des w_i à ces points fixes, on obtient encore des points de A. C'est ce procédé qui sera utilisé dans l'algorithme de la section 3.6.1.

3.4.3 Trouver un SIF : le théorème du collage

Le théorème de Hutchinson associe un attracteur à un SIF donné. Pour les applications, on peut aussi se poser le problème inverse : à partir d'une image de feuille, d'arbre ou de texture, comment trouver un SIF dont l'attracteur s'en approche ? Le théorème du collage, présenté par Michael Barnsley au milieu des années 1980 [Bar93], fournit une estimation utile pour cette recherche.

Théorème 3.4 **Théorème du collage**

Soit L un compact non vide du plan, et soit $\{w_1, \ldots, w_N\}$ un SIF de rapport de contraction k, d'attracteur A et d'opérateur de Hutchinson W, de sorte que $W(L) = w_1(L) \cup \cdots \cup w_N(L)$. Alors

$$d_H(L, A) \leq \frac{1}{1 - k}\, d_H(L, W(L)).$$

Démonstration. Il suffit d'appliquer le point (iii) du théorème 3.3, avec $B = L$ et $n = 0$. □

L'interprétation est la suivante : si on recouvre L, aussi fidèlement que possible, par des copies réduites et déformées de lui-même, $w_1(L), \ldots, w_N(L)$, c'est-à-dire à réaliser un *collage* de L à l'aide de ses propres photocopies réduites, alors l'attracteur du SIF ainsi obtenu ressemblera à L. Plus le collage est fidèle (plus $d_H(L, W(L))$ est petit), plus

l'estimation de l'écart est petite. Elle devient toutefois moins précise lorsque le rapport k s'approche de 1, à cause du facteur $1/(1-k)$.

Reste à trouver chaque w_i. Nous pouvons utiliser pour cela un résultat du chapitre 1 : par le théorème 1.15, une application affine est entièrement déterminée par les images de trois points non alignés. Il suffit donc de repérer trois points caractéristiques de la figure L (par exemple la base de la tige, la pointe et l'extrémité d'une feuille) et leurs positions sur la pièce L_i qu'on veut obtenir, puis de résoudre les deux systèmes 3×3 de l'équation (1.5).

Exemple 3.7 **Un arbre**

Cherchons un SIF dont l'attracteur ressemble à un arbre. Représentons un arbre stylisé par un tronc et une couronne dont les branches ressemblent à des arbres plus petits. On recouvre donc l'arbre par quatre pièces (figure 3.10) :

— le tronc, image de l'arbre entier par $w_1 \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0{,}03 & 0 \\ 0 & 0{,}42 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$, qui le réduit presque à un mince segment vertical ;

— deux grosses branches, images de l'arbre par des similitudes de rapport $0{,}58$, tournées de $38°$ vers la gauche et vers la droite, et posées au sommet du tronc : $w_{2,3} \begin{pmatrix} x \\ y \end{pmatrix} = 0{,}58\, r_{\pm 38°} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 0 \\ 0{,}42 \end{pmatrix}$;

— une cime, image de l'arbre par une homothétie de rapport $0{,}42$, posée un peu plus haut : $w_4 \begin{pmatrix} x \\ y \end{pmatrix} = 0{,}42 \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 0 \\ 0{,}6 \end{pmatrix}$.

Le SIF $\{w_1, w_2, w_3, w_4\}$ a pour rapport de contraction $0{,}58$; son attracteur est représenté à la figure 3.10, chaque pièce $w_i(A)$ dans sa couleur. On remarque que w_1 n'est pas une similitude : l'attracteur d'un SIF n'a pas à être autosimilaire au sens de la définition 3.1.

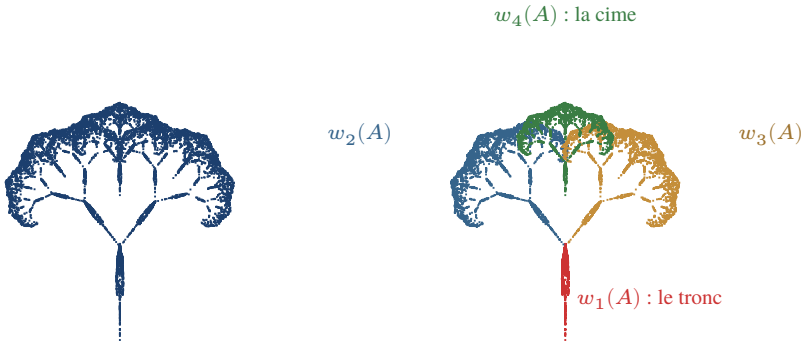


FIGURE 3.10 : L'attracteur du SIF de l'exemple 3.7 (à gauche) et sa décomposition en quatre pièces (à droite) : chaque point est coloré selon la dernière application utilisée pour l'obtenir. Les deux branches et la cime sont des copies réduites et tournées de l'arbre entier ; le tronc correspond à une copie fortement comprimée horizontalement.

Exemple 3.8 **La fougère de Barnsley**

Un exemple connu de cette méthode est la fougère de Barnsley, publiée en 1988 [Bar93]. Elle est l'attracteur du SIF de quatre applications affines $w_i \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a_i & b_i \\ c_i & d_i \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e_i \\ f_i \end{pmatrix}$ dont les coefficients sont donnés dans le tableau suivant (la colonne p_i servira à la section 3.6.2) :

i	a_i	b_i	c_i	d_i	e_i	f_i	p_i
1	0	0	0	0,16	0	0	0,01
2	0,85	0,04	−0,04	0,85	0	1,6	0,85
3	0,20	−0,26	0,23	0,22	0	1,6	0,07
4	−0,15	0,28	0,26	0,24	0	0,44	0,07

Examinons l'effet de chaque application (figure 3.11). La première écrase toute la fougère sur un petit segment vertical : c'est la base de la tige. Sa matrice est de déterminant nul ; c'est une application affine non inversible, comme celles de la section 1.6.8, ce que la définition 3.3 n'interdit pas. La deuxième, une similitude de rapport $\sqrt{0{,}85^2 + 0{,}04^2} \approx 0{,}851$, qui tourne légèrement le plan (d'environ 3° dans le sens horaire), envoie la fougère entière sur toute la fougère privée de sa première paire de folioles : c'est elle qui crée l'autosimilarité le long de la tige. Les deux dernières applications donnent les deux folioles du bas, à gauche et à droite ; la quatrième a un déterminant négatif, elle retourne la fougère, ce qui reproduit la disposition des folioles de part et d'autre de la tige.

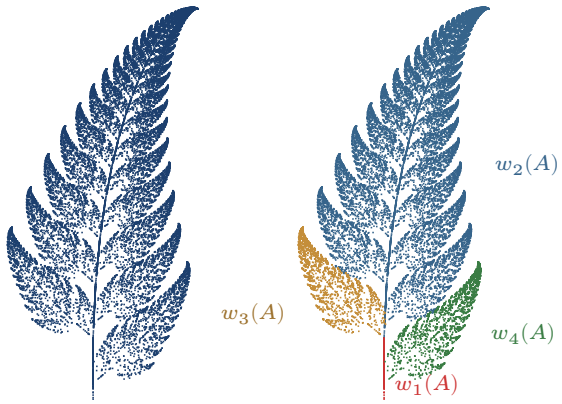


FIGURE 3.11 : La fougère de Barnsley (à gauche), attracteur d'un SIF de quatre applications affines, décrites par vingt-quatre coefficients. À droite, les quatre pièces $w_i(A)$: la base de la tige (en rouge), la fougère privée de ses folioles du bas (en bleu), et les deux folioles du bas (en or et en vert), qui sont elles-mêmes de petites fougères.

Remarque 3.7 **La compression fractale d'images**

Au début du chapitre 1, nous avons mentionné la possibilité de représenter une image par quelques transformations affines. La fougère de la figure 3.11 en est l'illustration : vingt-quatre nombres suffisent à la décrire, alors qu'une image matricielle de la même fougère en demande des centaines de milliers. À la fin des années 1980, Barnsley et son étudiant Arnaud Jacquin ont tiré de cette idée une méthode de *compression fractale* des images : l'image est découpée en petits blocs, et on cherche, pour chacun, une application affine qui l'obtient à partir d'un bloc plus grand de la même image, suivant le principe du théorème du collage. La méthode a été utilisée commercialement, notamment dans l'encyclopédie *Encarta* de Microsoft, avant d'être supplantée par la norme JPEG [Kom97].

3.5 La dimension de similitude généralisée

La définition 3.2 exigeait que toutes les copies aient le même rapport. Les SIF formés de similitudes de rapports différents donnent aussi des ensembles autosimilaires, comme ceux de la figure 3.12. L'argument heuristique de la section 3.3.2 peut être repris dans ce cas : si $E = w_1(E) \cup \cdots \cup w_N(E)$, où w_i est une similitude de rapport r_i, et si les copies ne se chevauchent pas, la « taille d-dimensionnelle » de E doit vérifier

$$M_d(E) = r_1^d\, M_d(E) + r_2^d\, M_d(E) + \cdots + r_N^d\, M_d(E),$$

ce qui n'est possible, avec une taille finie et non nulle, que si $r_1^d + \cdots + r_N^d = 1$.

Proposition 3.5 **L'équation de Moran**

Soient $N \geq 2$ et $r_1, \ldots, r_N \in]0,1[$. L'équation

$$r_1^d + r_2^d + \cdots + r_N^d = 1 \tag{3.1}$$

possède une unique solution $d > 0$.

Démonstration. Posons $\varphi(d) = r_1^d + \cdots + r_N^d$. Chaque terme $r_i^d = \mathrm{e}^{d \ln r_i}$ est une fonction continue et strictement décroissante de d, puisque $\ln r_i < 0$; la fonction φ est donc continue et strictement décroissante. De plus, $\varphi(0) = N \geq 2 > 1$ et $\varphi(d) \to 0$ lorsque $d \to \infty$. Par le théorème des valeurs intermédiaires, φ prend la valeur 1 ; sa stricte décroissance assure que cette valeur est atteinte une seule fois. □

Outil 3.2 **Résoudre l'équation de Moran**

L'équation (3.1) ne se résout généralement pas algébriquement, mais elle se résout bien graphiquement. Dans Desmos, on trace $y = 2(1/2)^x + 3(1/3)^x$ et $y = 1$: l'abscisse du point d'intersection, obtenue d'un clic, est la dimension cherchée (exercice 3.7). Le graphe illustre aussi la démonstration précédente : la courbe est décroissante, part de N en $x = 0$ et tend vers 0. Dans WolframAlpha, la requête `solve 2(1/2)^d + 3(1/3)^d = 1` donne directement une valeur approchée. Pour des élèves de cinquième secondaire, comparer cette résolution graphique avec la résolution par logarithmes du cas $Nr^d = 1$ (section 3.3) montre ce que les logarithmes permettent de faire, et ce qu'ils ne permettent pas.

Lorsque tous les rapports sont égaux à r, l'équation (3.1) devient $Nr^d = 1$, et on retrouve la définition 3.2. Il reste à préciser ce que veut dire « ne pas se chevaucher ». Les copies peuvent avoir des points communs sur leur frontière, comme les trois copies du triangle de Sierpiński, qui se touchent en des sommets ; on veut cependant exclure un recouvrement de leurs parties intérieures. La condition suivante, due à Moran [Mor46], rend cette idée précise.

Définition 3.4 **Condition de l'ouvert**

On dit que le SIF $\{w_1, \dots, w_N\}$ satisfait la **condition de l'ouvert** s'il existe un ensemble ouvert borné non vide V tel que

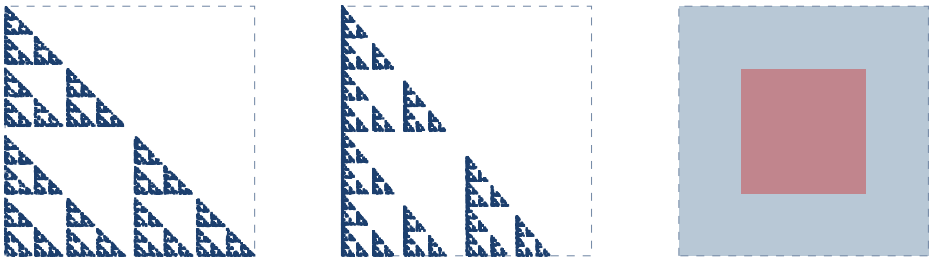
(a) $w_i(V) \subseteq V$ pour tout i ;

(b) les ensembles $w_1(V), \dots, w_N(V)$ sont deux à deux disjoints.

Pour le triangle de Sierpiński, on peut prendre pour V l'intérieur du triangle de départ : ses trois images sont les intérieurs des trois triangles de S_1, qui sont disjoints, parce que les triangles ne se touchent que par leurs sommets, qui sont sur leurs frontières. La distinction entre un ouvert et sa frontière, étudiée au chapitre 2, intervient donc directement : les copies peuvent avoir des sommets communs tout en ayant des intérieurs disjoints.

Définition 3.5 **Dimension de similitude généralisée**

Soit E l'attracteur d'un SIF formé de similitudes $w_1, \dots, w_N$ de rapports $r_1, \dots, r_N$, qui satisfait la condition de l'ouvert. On appelle **dimension de similitude généralisée** de E l'unique solution d de l'équation (3.1).



(a) trois copies de rapport 0,48 : disjointes

(b) rapports $\frac{1}{2}$, 0,4 et $\frac{1}{2}$: elles se touchent

(c) cinq copies de rapport $\frac{1}{2}$: elles se chevauchent

FIGURE 3.12 : Comparaison de trois SIF de similitudes. Le carré unité ouvert est indiqué en pointillé et noté V. En (a) et en (b), les images de V sont contenues dans V et deux à deux disjointes : la condition de l'ouvert est satisfaite, et on a représenté l'attracteur. En (c), on a représenté les cinq images du carré (quatre coins et un carré central) : la cinquième chevauche les quatre autres, et l'attracteur est le carré plein tout entier.

Exemple 3.9 **Pourquoi la condition de l'ouvert est nécessaire**

Considérons le SIF formé des cinq homothéties de rapport $\frac{1}{2}$ de la figure 3.12 (c), qui envoient le carré unité C sur ses quatre quarts et sur le carré de côté $\frac{1}{2}$ centré en $\left(\frac{1}{2}, \frac{1}{2}\right)$. Les quatre coins suffisent déjà à reconstituer C, et la cinquième copie est contenue dans C : on a donc $W(C) = C$, et l'attracteur est le carré plein, de dimension 2. L'équation de Moran, elle, donne $5\left(\frac{1}{2}\right)^d = 1$, c'est-à-dire $d = \ln 5 / \ln 2 \approx 2{,}32$. Cette valeur supérieure à 2 ne peut représenter la dimension de l'ensemble plan considéré. Le chevauchement de la cinquième copie conduit en effet à compter plusieurs fois certaines parties du carré.

Exemple 3.10 **Le calcul numérique d'une dimension**

L'attracteur de la figure 3.12 (b) est formé de deux copies de rapport $\frac{1}{2}$ et d'une copie de rapport $0{,}4$, qui se touchent sans se chevaucher (on prend pour V le carré unité ouvert). Sa dimension de similitude généralisée est la solution de

$$f(d) = 2\left(\frac{1}{2}\right)^d + (0{,}4)^d - 1 = 0.$$

On ne peut pas isoler d par les manipulations algébriques usuelles. Utilisons donc la méthode de Newton de la section 2.7.1 pour en obtenir une approximation. Avec $f'(d) = 2\left(\frac{1}{2}\right)^d \ln\frac{1}{2} + (0{,}4)^d \ln(0{,}4)$ et $d_0 = 1$, l'itération $d_{n+1} = d_n - f(d_n)/f'(d_n)$ donne :

n	d_n	$f(d_n)$
0	1,000 000 000	$4{,}0 \times 10^{-1}$
1	1,377 478 332	$5{,}3 \times 10^{-2}$
2	1,444 094 153	$1{,}3 \times 10^{-3}$
3	1,445 852 445	$8{,}9 \times 10^{-7}$
4	1,445 853 630	$4{,}0 \times 10^{-13}$

La dimension cherchée est donc $d \approx 1{,}445\,85$. Elle est comprise entre celle de trois copies de rapport $0{,}4$, $\ln 3 / \ln(2{,}5) \approx 1{,}199$, et celle de trois copies de rapport $\frac{1}{2}$, $\ln 3 / \ln 2 \approx 1{,}585$, comme on pouvait s'y attendre.

Remarque 3.8 **Et pour les SIF qui ne sont pas des similitudes ?**

La définition 3.5 ne s'applique qu'aux SIF formés de similitudes. Pour l'arbre de l'exemple 3.7, pour la fougère, ou pour tout attracteur d'un SIF d'applications affines quelconques, l'argument de la section 3.3.2 ne s'applique plus : une application qui écrase davantage dans une direction que dans l'autre ne multiplie pas une « taille d-dimensionnelle » par un facteur bien défini. Le calcul de la dimension de tels ensembles, qu'on dit *auto-affines*, demande d'autres outils, et fait encore aujourd'hui l'objet de recherches actives [Fal14].

3.6 Dessiner un attracteur à l'ordinateur

Après avoir établi l'existence de l'attracteur, nous allons voir comment en produire une image. Les figures fractales de ce chapitre ont été obtenues par deux algorithmes. Le

premier applique directement l'opérateur de Hutchinson ; le second utilise des choix aléatoires et demande généralement moins de calculs. Nous présenterons chacun en pseudo-code, puis en Python, avec les conventions de l'annexe A. Un nuage de points est représenté par un tableau $2 \times n$, dont chaque colonne contient les coordonnées d'un point, et une application affine par un couple $(A, \mathbf{t})$.

3.6.1 L'algorithme déterministe

Le point (ii) du théorème 3.3 suggère la méthode suivante : partir d'une figure B quelconque et lui appliquer W un nombre suffisant de fois. À l'ordinateur, la figure la plus simple est un ensemble fini de points ; on peut même se limiter à un point, puisqu'un singleton est un compact non vide.

Algorithme 3.1 **Algorithme déterministe**

Données : un SIF $\{w_1, \dots, w_N\}$; un nombre d'étapes n.

1. Choisir un point de départ P_0, et poser $\mathcal{P} = \{P_0\}$.
2. Répéter n fois : remplacer $\mathcal{P}$ par $W(\mathcal{P}) = w_1(\mathcal{P}) \cup \cdots \cup w_N(\mathcal{P})$.
3. Dessiner les points de $\mathcal{P}$.

Programme 3.1 : L'algorithme déterministe, appliqué au triangle de Sierpiński.

```
import numpy as np
import matplotlib.pyplot as plt

def applique(A, t, pts):
    """A : matrice 2x2, t : vecteur (2,), pts : tableau 2 x n."""
    A = np.asarray(A, dtype=float)
    t = np.asarray(t, dtype=float).reshape(2, 1)
    return A @ np.asarray(pts, dtype=float) + t

# Le SIF du triangle de Sierpinski : trois homothéties de rapport 1/2,
# centrées aux sommets S1, S2, S3 (partie translation : S_i / 2).
M = [[0.5, 0], [0, 0.5]]
SIF = [(M, [0, 0]), (M, [0.5, 0]), (M, [0.25, np.sqrt(3)/4])]

def W(pts):
    """Opérateur de Hutchinson : réunion des images par chaque w_i."""
    return np.hstack([applique(A, t, pts) for A, t in SIF])

pts = np.array([[0.], [0.]])      # un seul point de départ
for n in range(8):
    pts = W(pts)                  # 3^8 = 6561 points à la fin
plt.plot(pts[0], pts[1], ",", color="navy")
plt.axis("equal"); plt.axis("off"); plt.show()
```

Le point de départ choisi, l'origine, est le point fixe de w_1 ; par la remarque 3.6, il appartient à l'attracteur, et tous les points calculés aussi. Le dessin est donc constitué de 3^8 points appartenant tous à S. Il approche l'ensemble par un échantillon de ses propres points. Ces points sont de plus répartis sur tout le triangle : chaque point de S se trouve à une distance d'au plus $\left(\frac{1}{2}\right)^8 \approx 0{,}004$ de l'un des points calculés. En effet, W^8 est une contraction de rapport $\left(\frac{1}{2}\right)^8$ (théorème 3.1, appliqué huit fois), et $W^8(S) = S$; comme tous les points de S sont à distance au plus 1 de l'origine, $d_H(W^8(\{O\}), S) = d_H(W^8(\{O\}), W^8(S)) \leq$

$(\frac{1}{2})^8 d_H(\{O\}, S) \leq (\frac{1}{2})^8$.

La principale limite de cet algorithme est le nombre de points à calculer. L'étape n compte N^n points, et le nombre d'étapes nécessaires dépend du rapport de contraction k. Pour le triangle de Sierpiński, $N = 3$ et $k = \frac{1}{2}$, et une dizaine d'étapes suffisent. Pour la fougère de Barnsley, en revanche, $N = 4$ et la deuxième application a un rapport voisin de 0,85 : pour que $0{,}85^n$ devienne inférieur à un millième, il faut environ $n = 40$ étapes, soit $4^{40} \approx 1{,}2 \times 10^{24}$ points. Un tel nombre de points rend le calcul impraticable. L'algorithme aléatoire présenté ci-dessous permet d'éviter cette difficulté.

3.6.2 Le jeu du chaos

Le procédé peut d'abord être présenté comme un jeu en classe, à l'aide d'une feuille, d'une règle et d'un dé. On marque les trois sommets A, B et C d'un triangle équilatéral, et un point de départ P_0 quelconque. On lance le dé : sur 1 ou 2, on choisit le sommet A ; sur 3 ou 4, le sommet B ; sur 5 ou 6, le sommet C. On marque alors le point P_1 situé *à mi-chemin* entre P_0 et le sommet choisi. On relance le dé, et on marque P_2 à mi-chemin entre P_1 et le nouveau sommet choisi, et ainsi de suite (figure 3.13).

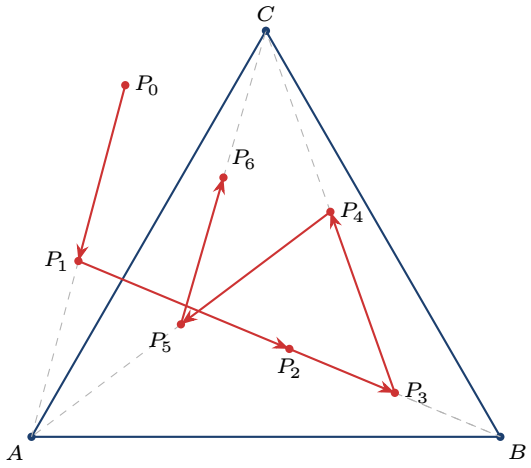


FIGURE 3.13 : Les six premières étapes du jeu du chaos, avec le point de départ P_0 (placé ici hors du triangle, ce qui est permis) et la suite de sommets A, B, B, C, A, C. Chaque flèche rouge mène au milieu du segment (en pointillé) qui joint le point courant au sommet choisi.

Bien que les sommets soient choisis au hasard, on voit apparaître une figure déterminée : après quelques milliers de coups, les points dessinent le triangle de Sierpiński (figure 3.14). Pour comprendre ce résultat, reprenons la définition de l'homothétie. Avancer à mi-chemin vers le sommet S_i, c'est appliquer au point courant l'homothétie de centre S_i et de rapport $\frac{1}{2}$, puisque $h_{S_i,1/2}(P) = \frac{1}{2}P + \frac{1}{2}S_i$ est exactement le milieu de P et de S_i (proposition 1.5). Or ces trois homothéties sont les trois applications w_1, w_2, w_3 du SIF du triangle de Sierpiński (exemple 3.3). Le jeu du chaos consiste donc à appliquer, à chaque coup, *une* des applications du SIF, choisie au hasard, plutôt que toutes à la fois comme dans l'algorithme déterministe.

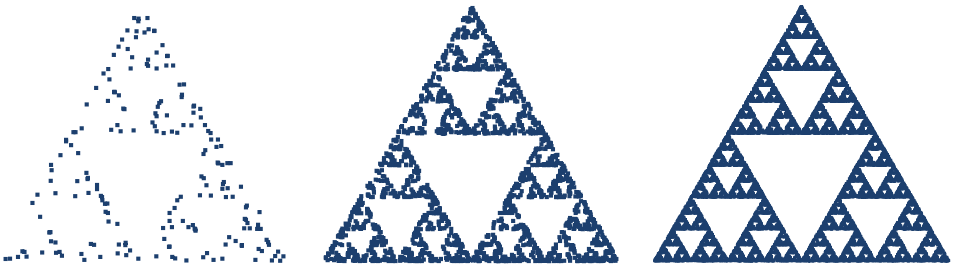


FIGURE 3.14 : Les points obtenus par le jeu du chaos pour le triangle de Sierpiński après 200, 2000 et 15 000 coups. Les vingt premiers points ont été omis.

Proposition 3.6 **Le jeu du chaos se rapproche de l'attracteur**

Soit A l'attracteur d'un SIF $\{w_1, \dots, w_N\}$ de rapport de contraction k, et soit P_0 un point quelconque du plan. Si $P_{m+1} = w_{i_m}(P_m)$ pour tout m, où les indices $i_0, i_1, \dots$ sont choisis de façon arbitraire, alors

$$d(P_m, A) \leq k^m \, d(P_0, A) \qquad \text{pour tout } m \in \mathbb{N},$$

où $d(P, A) = \min_{a \in A} d_E(P, a)$ désigne la distance d'un point à un ensemble.

Démonstration. Montrons d'abord que $d(w_i(P), A) \leq k \, d(P, A)$ pour tout point P et tout i ; le résultat s'obtient ensuite par récurrence sur m. Soit $a \in A$ un point de A le plus proche de P, de sorte que $d_E(P, a) = d(P, A)$. Puisque $w_i(a) \in w_i(A) \subseteq W(A) = A$,

$$d\big(w_i(P), A\big) \leq d_E\big(w_i(P), w_i(a)\big) \leq k \, d_E(P, a) = k \, d(P, A). \qquad \square$$

La distance à l'attracteur est donc divisée au moins par deux à chaque coup du jeu du chaos pour le triangle de Sierpiński, quels que soient les sommets tirés : après vingt coups, elle a été multipliée par $2^{-20} \approx 10^{-6}$, et le point est, à l'échelle d'une page, indiscernable d'un point de S. On omet donc les premiers points du dessin afin de réduire l'influence du point de départ.

La proposition 3.6 explique pourquoi les points se rapprochent de l'attracteur ; elle n'explique pas pourquoi ils le recouvrent *en entier*. Pour comprendre la répartition des points, il faut aussi tenir compte des probabilités. L'attracteur est la réunion des N^m petites pièces $w_{j_1} \circ w_{j_2} \circ \cdots \circ w_{j_m}(A)$, de taille au plus k^m fois celle de A. Chaque fois que les m derniers tirages ont été $j_m, \dots, j_2, j_1$, dans cet ordre, le point courant tombe (à une erreur négligeable près) dans la pièce correspondante. Or, à chaque coup, cette suite de m tirages a une probabilité strictement positive de se produire ; sur un grand nombre de coups, elle se produira donc, et même régulièrement. Toutes les petites pièces de l'attracteur finissent ainsi par être visitées, et le nuage de points le recouvre, aussi finement qu'on le souhaite.

Algorithme 3.2 **Jeu du chaos**

Données : un SIF $\{w_1, \dots, w_N\}$; des probabilités $p_1, \dots, p_N > 0$ de somme 1 ; un nombre de points n.

1. Choisir un point de départ P quelconque.
2. Répéter $n + 20$ fois :
   (a) tirer un indice i au hasard, la valeur i ayant la probabilité p_i d'être choisie ;
   (b) remplacer P par $w_i(P)$;
   (c) si plus de vingt tirages ont déjà eu lieu, dessiner P.

Le coût de cet algorithme est proportionnel au nombre n de points dessinés, et non plus exponentiel en le nombre d'étapes : quelques dizaines de milliers de calculs suffisent pour représenter la fougère, ce qui peut être effectué en une fraction de seconde.

Programme 3.2 : Le jeu du chaos, appliqué à la fougère de Barnsley. La fonction applique est celle du programme précédent.

```
rng = np.random.default_rng()

def jeu_du_chaos(SIF, probas, n=50_000, rejet=20):
    """n points P_(k+1) = w_i(P_k), avec w_i tirée au hasard."""
    p = np.zeros((2, 1))                    # départ : l'origine
    pts = np.empty((2, n))
    choix = rng.choice(len(SIF), size=n + rejet, p=probas)
    for k, i in enumerate(choix):
        A, t = SIF[i]
        p = applique(A, t, p)
        if k >= rejet:                      # on oublie les premiers
            pts[:, k - rejet] = p[:, 0]
    return pts

# La fougère de Barnsley : une ligne (A, t) par application.
FOUGERE = [([[ 0.00,  0.00], [ 0.00, 0.16]], [0, 0.00]),
           ([[ 0.85,  0.04], [-0.04, 0.85]], [0, 1.60]),
           ([[ 0.20, -0.26], [ 0.23, 0.22]], [0, 1.60]),
           ([[-0.15,  0.28], [ 0.26, 0.24]], [0, 0.44])]
pts = jeu_du_chaos(FOUGERE, [0.01, 0.85, 0.07, 0.07])
plt.plot(pts[0], pts[1], ",", color="darkgreen")
plt.axis("equal"); plt.axis("off"); plt.show()
```

3.6.3 Choisir les probabilités

Les probabilités p_i n'ont aucune influence sur la figure limite : tant qu'elles sont toutes strictement positives, l'argument précédent garantit que le nuage finit par recouvrir tout l'attracteur. Elles ont en revanche une grande influence sur la *vitesse* à laquelle il le recouvre. Pour le triangle de Sierpiński, les trois pièces sont de même taille, et le choix $p_1 = p_2 = p_3 = \frac{1}{3}$ est naturel. Pour la fougère, un tirage uniforme ($p_i = \frac{1}{4}$) donnerait une répartition peu efficace : un point sur quatre tomberait sur la minuscule base de la tige, tandis que les folioles, qui occupent presque toute l'image, resteraient clairsemées.

Une façon de choisir les probabilités consiste à prendre p_i proportionnelle à la proportion de l'aire que la pièce $w_i(A)$ occupe dans l'image. Par la proposition 1.4, une application

affine de partie linéaire A_i multiplie les aires par $|\det(A_i)|$; on pose donc

$$p_i \approx \frac{|\det(A_i)|}{|\det(A_1)| + \cdots + |\det(A_N)|},$$

en remplaçant un déterminant nul par une petite valeur positive, pour que la pièce correspondante soit tout de même visitée. Pour la fougère, les déterminants valent 0, 0,7241, 0,1038 et −0,1088, ce qui donne des probabilités voisines de 0,77, 0,11 et 0,12 pour les trois dernières pièces : ce sont, à peu de chose près, les valeurs 0,85, 0,07 et 0,07 retenues par Barnsley, qui avait réservé 0,01 à la tige. Un SIF muni de telles probabilités s'appelle un **SIF avec probabilités** ; Barnsley et Demko ont montré qu'il décrit, en plus de la figure, une répartition des points sur celle-ci [BD85], un point de vue utile pour représenter des images en niveaux de gris.

Pour la classe 3.4 **Le jeu du chaos en classe**

Le jeu du chaos permet de réunir trois sujets étudiés au secondaire : la géométrie (le point milieu, l'homothétie), les probabilités (un dé équilibré, des événements équiprobables) et la notion de limite. Chaque élève joue une vingtaine de coups sur un transparent où le même triangle est tracé ; on omet les cinq premiers points, puis on superpose les transparents de toute la classe. Le triangle de Sierpiński apparaît, ce qui peut surprendre les élèves qui avaient prévu une répartition uniforme dans le triangle. La question « pourquoi n'y a-t-il jamais de point dans le grand trou central ? » admet une réponse accessible : le milieu d'un segment qui joint un point du triangle à l'un de ses sommets tombe toujours dans le petit triangle de coin associé à ce sommet, jamais dans le trou central. Une variante avec les quatre sommets d'un carré, qui conduit à un autre résultat, fait l'objet de l'exercice 3.11.

3.7 Des monstres aux modèles

Pour terminer, revenons sur deux moments de l'histoire des fractales : la construction de contre-exemples en analyse au XIXe siècle, puis l'utilisation de ces objets pour modéliser des phénomènes dans la seconde moitié du XXe siècle.

3.7.1 La fonction de Weierstrass

Jusqu'au milieu du XIXe siècle, la plupart des mathématiciens tenaient pour acquis qu'une fonction continue est dérivable, sauf peut-être en quelques points isolés, comme la valeur absolue en 0. Cette idée correspond à une intuition familière : le graphe d'une fonction continue se trace « sans lever le crayon », et un trait de crayon a, presque partout, une direction. Le 18 juillet 1872, devant l'Académie royale des sciences de Prusse, Karl Weierstrass présente un contre-exemple à cette intuition avec la fonction

$$W(x) = \sum_{n=0}^{\infty} a^n \cos\left(b^n \pi x\right),$$

où $0 < a < 1$ et où b est un entier impair assez grand. Chaque terme est une fonction lisse ; la série converge, car ses termes sont majorés en valeur absolue par ceux de la série géométrique $\sum a^n$; et sa somme est une fonction continue. Pourtant, Weierstrass démontre qu'elle n'est dérivable en *aucun* point[1]. Pour interpréter ce comportement, observons

1. Weierstrass exigeait $ab > 1 + \frac{3}{2}\pi$. Le mathématicien anglais G. H. Hardy a montré en 1916 que la condition $ab \geq 1$ suffit, ce qui autorise le choix $a = \frac{1}{2}$, $b = 3$ de la figure 3.15.

que les termes ajoutés ont des amplitudes a^n qui décroissent, mais des fréquences b^n qui croissent beaucoup plus vite : chaque terme ajoute au graphe des oscillations plus petites, mais proportionnellement plus raides, et la pente ne se stabilise jamais.

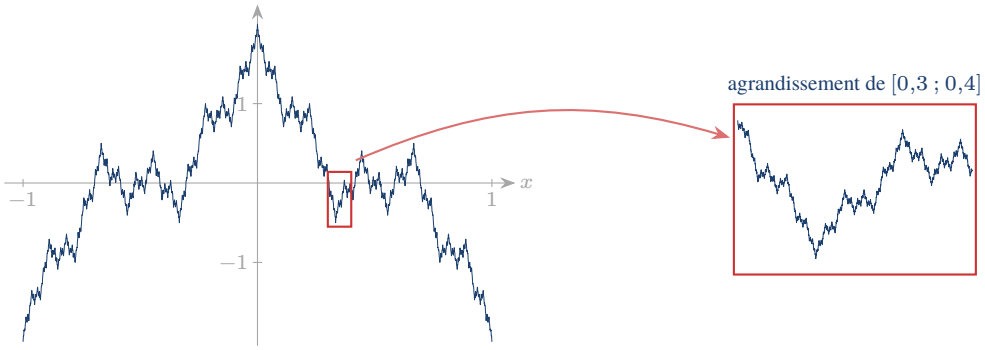


FIGURE 3.15 : Le graphe de la fonction de Weierstrass $W(x) = \sum a^n \cos(b^n \pi x)$ pour $a = \frac{1}{2}$ et $b = 3$ (somme des dix-huit premiers termes), et un agrandissement de la portion encadrée. L'agrandissement n'est pas une copie exacte du graphe entier, mais il présente des irrégularités comparables : aucune échelle ne laisse apparaître de tangente.

La fonction de Weierstrass partage avec la courbe de von Koch l'irrégularité à toutes les échelles, et elle en partage aussi une forme d'autosimilarité. En isolant le premier terme de la série et en remarquant que les suivants forment encore une fonction de Weierstrass, on obtient en effet l'identité

$$W(x) = \cos(\pi x) + a\, W(bx).$$

Au terme lisse $\cos(\pi x)$ près, le graphe de W au-dessus d'un intervalle de longueur ℓ est donc une copie de son graphe au-dessus d'un intervalle de longueur $b\ell$, comprimée d'un facteur b horizontalement, mais seulement d'un facteur $1/a$ verticalement. Une telle copie, obtenue par une application affine qui n'est pas une similitude, est dite *auto-affine*. Il a été démontré en 2018, que le graphe de W est une fractale de dimension de Hausdorff $2 + \ln a / \ln b$, soit environ 1,37 pour la figure 3.15 [She18].

Ces fonctions ont suscité des réactions marquées chez plusieurs mathématiciens. Charles Hermite écrit en 1893 à son collègue Thomas Stieltjes : « Je me détourne avec effroi et horreur de cette plaie lamentable des fonctions continues qui n'ont point de dérivées. » Henri Poincaré, en 1908, exprime lui aussi des réserves :

> La logique parfois engendre des monstres. Depuis un demi-siècle on a vu surgir une foule de fonctions bizarres qui semblent s'efforcer de ressembler aussi peu que possible aux honnêtes fonctions qui servent à quelque chose. […] Autrefois, quand on inventait une fonction nouvelle, c'était en vue de quelque but pratique ; aujourd'hui, on les invente tout exprès pour mettre en défaut les raisonnements de nos pères, et on n'en tirera jamais que cela. [Poi08]

Les applications développées une soixantaine d'années plus tard allaient toutefois donner un autre rôle à ces objets : ils serviraient à représenter des côtes, des montagnes et des variations boursières.

Note historique 3.3 Karl Weierstrass



Karl Weierstrass (1815–1897), portrait de Conrad Fehr (1895). Domaine public, *Wikimedia Commons*.

Karl Weierstrass (1815–1897) est né en Westphalie. Son père, un fonctionnaire, l'envoie étudier le droit et l'administration publique à Bonn ; il y passe quatre ans, surtout à faire de l'escrime et à fréquenter les tavernes, et en repart sans diplôme. Il se tourne alors vers l'enseignement, et passe près de quinze ans comme enseignant au secondaire dans de petites villes de Prusse, à enseigner non seulement les mathématiques et la physique, mais aussi l'allemand, la botanique, la géographie, la gymnastique et la calligraphie. Il poursuit ses recherches seul, le soir, sans contact avec le milieu mathématique. Avant d'occuper une place majeure en analyse, Weierstrass a donc exercé pendant plusieurs années le métier auquel vous vous préparez.

En 1854, un article sur les fonctions abéliennes, publié dans une importante revue mathématique allemande, attire l'attention sur ses travaux : l'université de Königsberg lui décerne un doctorat honorifique, et il est appelé à Berlin en 1856, à quarante ans. Il y devient l'un des professeurs les plus influents de son siècle. Ses travaux ont contribué à préciser les exigences de rigueur de l'analyse moderne, et en particulier les définitions de la limite et de la continuité au moyen des ε et des δ, celles-là mêmes que nous avons utilisées au chapitre 2. Parmi ses élèves figurent Georg Cantor, Hermann Schwarz, Gösta Mittag-Leffler (le directeur de thèse de von Koch) et Sofia Kovalevskaïa, à qui il donna des leçons privées pendant quatre ans, puisque l'université de Berlin refusait alors les femmes ; elle deviendra en 1889, à Stockholm, l'une des premières femmes en Europe à occuper une chaire de professeure de mathématiques. Sa fonction continue nulle part dérivable, présentée en 1872, n'a été publiée qu'en 1875, par son collègue Paul du Bois-Reymond [ORh].

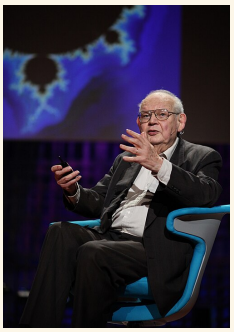
3.7.2 Benoît Mandelbrot et la géométrie de la nature

En 1967, Benoît Mandelbrot publie dans la revue *Science* un article intitulé : « Combien mesure la côte de la Grande-Bretagne ? » [Man67]. Il y reprend des mesures faites par le physicien Lewis Fry Richardson, qui avait remarqué que la longueur mesurée d'une côte dépend de l'unité de mesure : plus le compas avec lequel on la parcourt a une petite ouverture ε, plus il épouse de baies et de caps, et plus la longueur trouvée est grande. Richardson avait observé que cette longueur croît comme une puissance, $L(\varepsilon) \approx C\,\varepsilon^{1-D}$, avec $D \approx 1{,}25$ pour la côte ouest de la Grande-Bretagne. Mandelbrot y reconnaît une dimension fractionnaire : on retrouve un comportement analogue pour la courbe de von Koch, dont la longueur « mesurée » avec des segments de longueur $\left(\frac{1}{3}\right)^n$ vaut $\left(\frac{4}{3}\right)^n$, soit ε^{1-D} avec $D = \ln 4/\ln 3$. Une côte n'est pas autosimilaire au sens strict, mais elle l'est au sens statistique de la remarque 3.2, et sa dimension mesure son degré d'irrégularité.

Mandelbrot généralise ensuite cette idée à plusieurs phénomènes : les montagnes, les nuages, la répartition des galaxies, les turbulences, les réseaux de rivières. Son livre de 1982, *The Fractal Geometry of Nature* [Man82], s'ouvre sur une phrase restée célèbre : « Les nuages ne sont pas des sphères, les montagnes ne sont pas des cônes, les côtes ne sont pas des cercles, l'écorce n'est pas lisse, et l'éclair ne se propage pas en ligne droite » (notre traduction). C'est aussi lui qui a fait connaître, à partir de 1980, l'ensemble qui porte son nom, obtenu en itérant la fonction $z \mapsto z^2 + c$ dans le plan complexe, qui allait devenir l'une des représentations les plus connues de la géométrie fractale. L'étude de cet

ensemble, qui ne provient pas d'un SIF, dépasse toutefois le cadre de ce chapitre.

Note historique 3.4 Benoît Mandelbrot



Benoît Mandelbrot (1924–2010), en 2010. Photo : Steve Jurvetson, licence CC BY 2.0, *Wikimedia Commons*.

Benoît Mandelbrot (1924–2010) est né à Varsovie, dans une famille juive d'origine lituanienne qui émigre en France en 1936. Son oncle, Szolem Mandelbrojt, est professeur de mathématiques au Collège de France. Pendant l'Occupation, la famille survit en se cachant dans le centre de la France, et la scolarité de Mandelbrot est constamment interrompue ; il racontera plus tard avoir développé, faute de formation régulière, une approche des problèmes fondée sur la visualisation géométrique. Admis en 1944 à l'École normale supérieure et à l'École polytechnique, il choisit la seconde, étudie ensuite l'aéronautique au California Institute of Technology, puis soutient à Paris, en 1952, une thèse sur la distribution statistique des mots dans les textes.

En 1958, il choisit de poursuivre ses recherches chez IBM, près de New York, plutôt que dans une université. Il y restera trente-cinq ans et pourra s'intéresser à des domaines variés : linguistique, économie, hydrologie, physique, astronomie. Ces déplacements entre les disciplines suscitent parfois la méfiance des spécialistes. Ils lui permettent toutefois de reconnaître des structures communes dans des phénomènes étudiés séparément. Il forge le mot « fractal » en 1975, pour le titre de son livre *Les objets fractals : forme, hasard et dimension* [Man75]. Il reçoit plusieurs distinctions à la fin de sa vie (prix Wolf de physique en 1993, prix du Japon en 2003) et obtient en 1999, à 75 ans, un poste de professeur titulaire à l'université Yale. Il meurt à Cambridge, au Massachusetts, en 2010 [ORb].

3.7.3 Les fractales et la finance

Les applications financières des travaux de Mandelbrot ont fait l'objet de nombreuses discussions. Il s'est intéressé à ce domaine dès le début des années 1960, avant même d'introduire le mot « fractal ».

Le modèle classique. En 1900, dans sa thèse *Théorie de la spéculation* [Bac00], le mathématicien français Louis Bachelier propose de modéliser le cours d'une action comme une marche au hasard : chaque jour, le prix varie d'une quantité aléatoire, indépendante des variations précédentes et distribuée selon une loi normale. À la limite, on obtient ce que les physiciens appelleront le *mouvement brownien*. Ce modèle, oublié pendant un demi-siècle, devient dans les années 1950 à 1970 le fondement de la finance moderne : la théorie du portefeuille et la formule de Black et Scholes (1973), qui sert à fixer le prix des options, s'appuient sur ce cadre.

Ce que Mandelbrot observe. En 1963, Mandelbrot étudie les variations du prix du coton sur plusieurs décennies [Man63]. Il fait deux observations qui s'accordent mal avec le modèle de Bachelier.

— *Les grandes variations sont beaucoup trop fréquentes.* Dans une loi normale, un écart de plus de cinq écarts-types est si rare qu'on ne devrait pratiquement jamais l'observer ; dans les données réelles, de tels sauts se produisent régulièrement. Les distributions observées ont des « queues épaisses », que Mandelbrot propose de

décrire par les lois stables de Paul Lévy. Il appellera ce phénomène l'*effet Noé*, en souvenir du déluge : des changements brusques et de grande ampleur.

— *La forme des variations ne dépend pas de l'échelle de temps.* Après une simple mise à l'échelle, la distribution des variations quotidiennes a la même forme que celle des variations mensuelles. Autrement dit, le graphe du cours est statistiquement autosimilaire, ou plutôt auto-affine, puisque les échelles de temps et de prix ne se dilatent pas du même facteur : sans les étiquettes des axes, il est impossible de dire si un graphique boursier couvre une journée, un mois ou une année.

Mandelbrot relèvera plus tard un troisième fait : les périodes agitées, comme les périodes calmes, tendent à se prolonger et à se regrouper, ce qui suggère une forme de mémoire dans les variations. Il parle cette fois de l'*effet Joseph*, en souvenir des sept années d'abondance suivies de sept années de famine du récit biblique.

Un SIF pour représenter des variations boursières. Pour faire comprendre ses modèles à un large public, Mandelbrot a imaginé des « caricatures » (*cartoons*) de cours boursiers, construites selon un procédé analogue à celui de la courbe de von Koch [Man99]. On part d'une tendance, le segment joignant $(0,0)$ à $(1,1)$, et d'un *générateur*, une ligne brisée qui monte, redescend, puis remonte : $(0,0)$, $(\frac{4}{9},\frac{2}{3})$, $(\frac{5}{9},\frac{1}{3})$, $(1,1)$. On remplace ensuite chaque segment par une copie du générateur, ajustée à ses extrémités, et on recommence (figure 3.16). Dans le langage de ce chapitre, la courbe limite est l'attracteur du SIF formé des trois applications affines

$$w_1\begin{pmatrix}x\\y\end{pmatrix} = \begin{pmatrix}\frac{4}{9} & 0\\ 0 & \frac{2}{3}\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}, \qquad w_2\begin{pmatrix}x\\y\end{pmatrix} = \begin{pmatrix}\frac{1}{9} & 0\\ 0 & -\frac{1}{3}\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix} + \begin{pmatrix}\frac{4}{9}\\ \frac{2}{3}\end{pmatrix},$$
$$w_3\begin{pmatrix}x\\y\end{pmatrix} = \begin{pmatrix}\frac{4}{9} & 0\\ 0 & \frac{2}{3}\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix} + \begin{pmatrix}\frac{5}{9}\\ \frac{1}{3}\end{pmatrix},$$

qui ne sont pas des similitudes : elles compriment davantage le temps que les prix. Les rapports ont été choisis pour que chaque morceau vérifie $|\Delta y| = \sqrt{\Delta x} : \frac{2}{3} = \sqrt{4/9}$ et $\frac{1}{3} = \sqrt{1/9}$. On retrouve ainsi le comportement moyen des variations du mouvement brownien de Bachelier, dont l'amplitude sur une durée Δt est proportionnelle à $\sqrt{\Delta t}$.

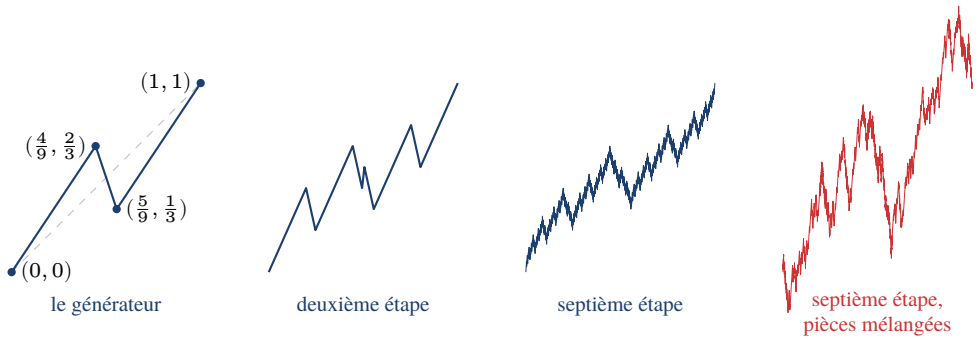


FIGURE 3.16 : Une « caricature » de cours boursier à la Mandelbrot. À chaque étape, chaque segment est remplacé par une copie du générateur (à gauche). La courbe obtenue (troisième panneau) est trop régulière pour ressembler à un vrai cours ; en tirant au hasard, pour chaque segment, l'ordre dans lequel on place les trois morceaux du générateur, on obtient une courbe (à droite, en rouge) dont l'allure rappelle celle d'un graphique boursier.

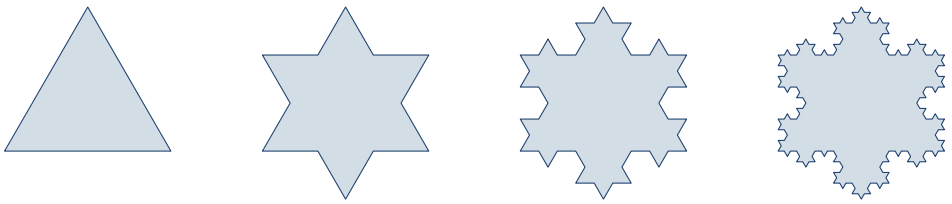
La version déterministe est trop régulière, mais il suffit, à chaque remplacement, de mélanger au hasard l'ordre des trois morceaux du générateur pour obtenir une courbe dont l'allure est plus proche de celle d'un cours boursier (figure 3.16, à droite). En modifiant le générateur, par exemple en déplaçant ses deux points intérieurs, Mandelbrot obtenait

des courbes où alternent périodes calmes et périodes de forte agitation, comme dans les marchés réels : c'est l'idée de son *modèle multifractal* des rendements financiers, développé en 1997 avec Laurent Calvet et Adlai Fisher.

Les conséquences. Mandelbrot souligne une limite des modèles utilisés en finance : les modèles gaussiens sous-estiment systématiquement le risque de catastrophe. Le 19 octobre 1987, l'indice Dow Jones a perdu près de 23 % de sa valeur en une seule séance, un événement auquel un modèle gaussien, calibré sur les variations habituelles, n'attribue qu'une probabilité extrêmement faible, incompatible avec la fréquence des événements observés. Dans *The (Mis)behavior of Markets*, écrit avec le journaliste Richard Hudson en 2004 [MH04], il plaide pour des modèles qui tiennent compte de les queues épaisses et la mémoire des marchés. Ces idées ont été reçues avec prudence : les modèles fractals sont plus difficiles à manipuler, et la formule de Black et Scholes reste un outil de base. La crise financière de 2008 a cependant ravivé l'intérêt pour ses travaux, et les « queues épaisses » font aujourd'hui partie du vocabulaire courant de la gestion des risques.

Exercices

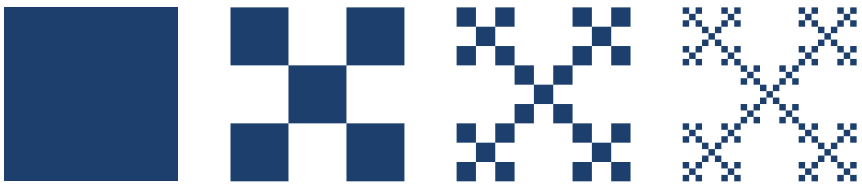
Exercice 3.1. Considérons la construction du flocon de von Koch à partir d'un triangle équilatéral de côté 1 : à chaque étape, on remplace chaque segment du contour par les quatre segments de la construction de la section 3.1.1, la pointe étant dirigée vers l'extérieur.



étape 0 étape 1 étape 2 étape 3

a) Montrer que le périmètre du flocon est infini.
b) Déterminer l'aire du flocon.

Exercice 3.2. On construit un flocon de neige à partir d'un carré plein de côté a, qui forme l'étape 1. À chacune des étapes suivantes, on découpe chaque carré en neuf carrés égaux, et on ne conserve que les quatre carrés de coin et le carré central.



étape 1 étape 2 étape 3 étape 4

a) Trouver l'aire A_n de la figure obtenue à l'étape n.
b) Déterminer la dimension de similitude du flocon de neige.

Exercice 3.3. Soit E l'ensemble des nombres réels de l'intervalle $[0, 1]$ qui possèdent une écriture décimale dont toutes les décimales sont paires.

a) Décrire une construction de l'ensemble E en débutant avec l'intervalle $[0, 1]$, duquel on retranche, à chaque étape, des intervalles.
b) Expliquer pourquoi E est autosimilaire, puis calculer sa dimension de similitude.

Exercice 3.4. Pour chacun des ensembles autosimilaires suivants, calculer l'aire et la dimension de similitude.

a) Le triangle de Sierpiński construit à partir d'un triangle équilatéral dont les côtés sont de longueur 1.
b) Le tapis de Sierpiński construit à partir d'un carré dont les côtés sont de longueur 1.

Exercice 3.5. Pour chacun des ensembles autosimilaires suivants, de l'espace à trois dimensions, calculer le volume et la dimension de similitude.

a) L'**éponge de Menger**, construite à partir d'un cube dont les arêtes sont de longueur 1 : on découpe le cube en 27 cubes égaux, on retire le cube central et les six cubes situés au centre des faces, puis on recommence dans chacun des cubes restants.
b) Le **tétraèdre de Sierpiński**, construit à partir d'un tétraèdre régulier dont les arêtes sont de longueur 1 : en joignant les milieux des arêtes, on ne conserve que les quatre tétraèdres situés aux sommets, puis on recommence dans chacun d'eux.

Exercice 3.6. Pour chacune des paires d'ensembles ci-dessous, trouver $d_H(A, B)$, où d_H est la distance de Hausdorff (définition 2.9).

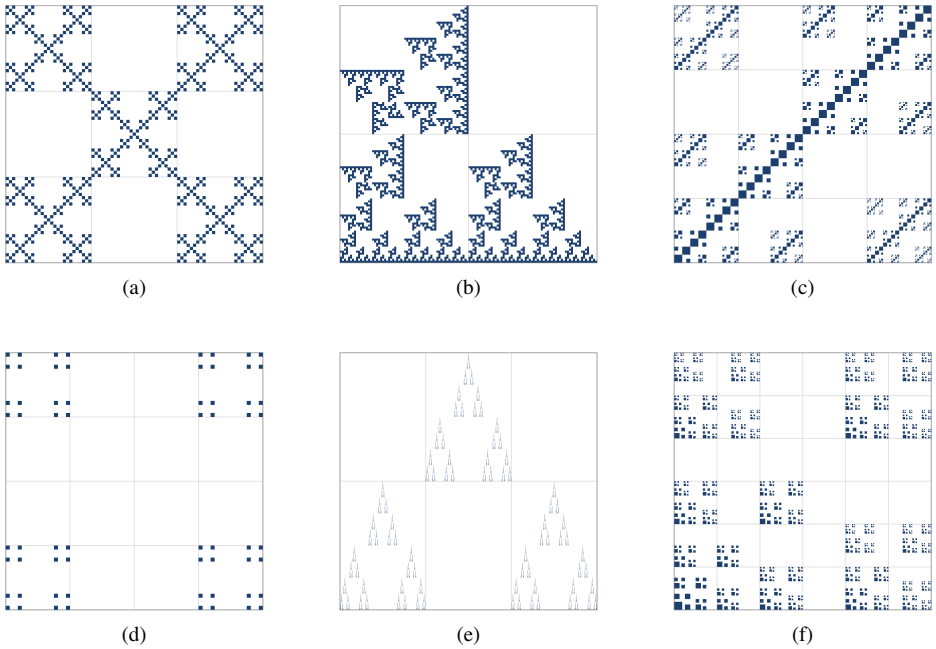
a) $A = \{(x_1, y_1)\}$ et $B = \{(x_2, y_2)\}$.
b) $A = \{(1, 1)\}$ et $B = \{(x, y) : x^2 + y^2 \leq r^2\}$.
c) $A = \{(x, 0) : -1 \leq x \leq 1\}$ et $B = \{(0, y) : -2 \leq y \leq 2\}$.
d) $A = \{(x, y) : (x - h_1)^2 + (y - k_1)^2 \leq r_1^2\}$ et $B = \{(x, y) : (x - h_2)^2 + (y - k_2)^2 \leq r_2^2\}$.

Exercice 3.7. Un SIF est formé de deux similitudes de rapport $\frac{1}{2}$ et de trois similitudes de rapport $\frac{1}{4}$, et satisfait la condition de l'ouvert.

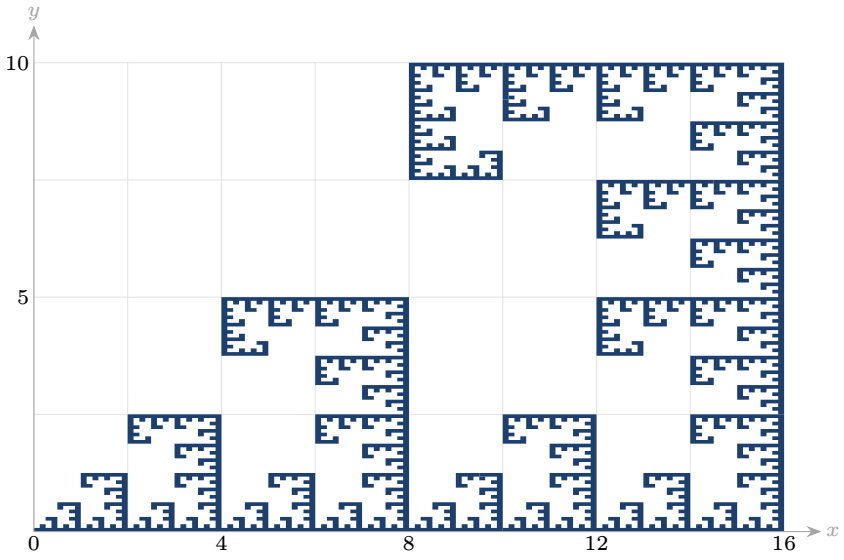
a) Quel est le rapport de contraction du SIF ?
b) Quelle est la dimension de similitude généralisée de l'attracteur du SIF ?
c) Si le SIF avait plutôt été formé de deux similitudes de rapport $\frac{1}{2}$ et de trois similitudes de rapport $\frac{1}{3}$, quelle aurait été la dimension de similitude généralisée de son attracteur ?

Exercice 3.8. Chacun des ensembles E ci-dessous, tracés dans le carré unité, est l'attracteur d'un SIF formé d'applications affines. Pour chacun :

a) trouver des applications affines $w_1, \ldots, w_n$ telles que E soit l'attracteur du SIF $\{w_1, \ldots, w_n\}$;
b) utiliser un ordinateur pour construire une image qui ressemble à E ;
c) trouver, si elle existe, la dimension de similitude de E ;
d) trouver, si elle existe, la dimension de similitude généralisée de E.



Exercice 3.9. Soit $\mathcal{S}$ le SIF dont l'attracteur A est représenté ci-dessous.



a) Trouver trois fonctions f_1, f_2 et f_3 telles que $\mathcal{S} = \{f_1, f_2, f_3\}$.
b) Quelle est la dimension de similitude généralisée de A ?

Exercice 3.10. Considérons le SIF formé des contractions

$$w_1\begin{pmatrix}x\\y\end{pmatrix} = \frac{1}{3}\begin{pmatrix}x\\y\end{pmatrix}, \qquad w_2\begin{pmatrix}x\\y\end{pmatrix} = \frac{1}{3}\begin{pmatrix}x\\y\end{pmatrix} + \begin{pmatrix}\frac{2}{3}\\0\end{pmatrix},$$

et soit $C = \{(x, y) : 0 \leq x \leq 1,\ 0 \leq y \leq 1\}$.

a) Représenter C dans le plan.
b) Trouver $W(C)$ et le représenter dans le plan.
c) Trouver $W^2(C)$ et le représenter dans le plan.
d) Que vaut $d_H(W(C), C)$?
e) Que vaut le rapport de contraction du SIF ?

f) Soit A l'attracteur du SIF. Utiliser le théorème de Hutchinson pour trouver une valeur de n telle que $d_H(W^n(C), A) \leq 0{,}01$.

Exercice 3.11. Considérons le jeu du chaos associé aux les quatre sommets $(0,0)$, $(1,0)$, $(1,1)$ et $(0,1)$ du carré unité, chacun étant choisi avec une probabilité de $\frac{1}{4}$.

a) Écrire le SIF correspondant si, à chaque coup, on avance à mi-chemin vers le sommet choisi. Quel est son attracteur ? Qu'observera-t-on à l'écran ?

b) Même question si, à chaque coup, on avance plutôt des deux tiers du chemin vers le sommet choisi. Calculer la dimension de similitude de l'attracteur.

Exercice 3.12. *(Défi facultatif.)* Sans consulter le tableau de l'exemple 3.8, répondre aux sous-questions de l'exercice 3.8 pour la fougère de Barnsley de la figure 3.11.

Exercice 3.13. Les fractales n'apparaissent pas explicitement dans le programme du secondaire, mais elles sont souvent présentées aux élèves comme activité d'enrichissement.

a) Dresser la liste des notions du programme du secondaire qui peuvent être réinvesties dans l'étude du flocon de von Koch (exercice 3.1) ou du triangle de Sierpiński. Préciser le niveau correspondant à chacune.

b) Le résultat « périmètre infini, aire finie » étonne les élèves. Comment le présenter sans disposer de la notion de limite ? Quelles précautions faut-il prendre ?

c) Quels sont les avantages et les limites d'une activité où les élèves jouent au jeu du chaos à la main, puis le programment ?

4 Les chaînes de Markov et l'algorithme PageRank de Google

Lorsqu'on tape quelques mots dans un moteur de recherche, des millions de pages contiennent souvent ces mots. Pour les présenter dans un ordre utile, le moteur doit décider lesquelles sont les plus *importantes*. À la fin des années 1990, deux étudiants de l'Université Stanford, Sergey Brin et Larry Page, ont proposé une réponse simple : imaginons un internaute qui navigue au hasard, en cliquant sur les liens qu'il rencontre, et mesurons la proportion du temps qu'il passe sur chaque page. Une page souvent visitée est une page importante. Cette idée est à l'origine de l'algorithme *PageRank* et du moteur de recherche Google.

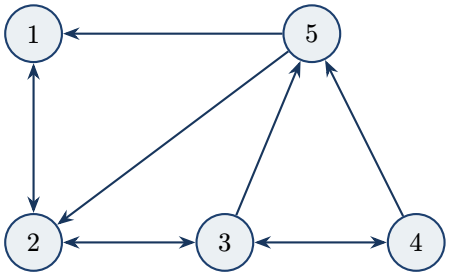


FIGURE 4.1 : Un *mini-web* de cinq pages. Une flèche de j vers i représente un lien, sur la page j, qui mène à la page i ; une double flèche représente deux liens, un dans chaque sens. Un internaute qui clique au hasard passera-t-il plus de temps sur la page 2, qui reçoit quatre liens, ou sur la page 1, qui n'en reçoit que deux ?

Le déplacement de cet internaute est un exemple de *chaîne de Markov* : un phénomène aléatoire qui évolue par étapes, et dont l'état à l'étape $n+1$ ne dépend que de son état à l'étape n, et non de tout son passé. Ce modèle, proposé par le mathématicien russe Andreï Markov au début du XXe siècle, sert aujourd'hui en météorologie, en démographie, en génétique, en finance, en reconnaissance de la parole et dans les modèles qui génèrent du texte.

Nous commencerons par quelques rappels sur les probabilités conditionnelles et les variables aléatoires discrètes. Nous étudierons ensuite deux modèles simples, un modèle météorologique à deux états et un modèle de migration entre la ville, la banlieue et la campagne. Ils nous permettront de définir les chaînes de Markov et leur matrice de transition. Nous verrons que le comportement à long terme d'une chaîne est décrit par un vecteur propre associé à la valeur propre 1, ce qui nous ramènera aux notions de l'annexe D, puis au théorème du point fixe de Banach du chapitre 2. Nous présenterons enfin l'algorithme PageRank, en tenant compte des difficultés que pose le Web réel : les pages sans liens sortants et les régions dont on ne peut plus sortir.

Note historique 4.1 Andreï Markov



Andreï Markov (1856–1922), vers 1875. Domaine public, *Wikimedia Commons*.

Andreï Andreïevitch Markov (1856–1922) est né à Riazan, en Russie. Enfant de santé fragile, atteint d'une malformation à la jambe, il dut se déplacer avec des béquilles jusqu'à l'âge d'une dizaine d'années. Il étudia à l'Université de Saint-Pétersbourg, où il fut l'élève de Pafnouti Tchebychev, dont il prolongea les travaux en théorie des probabilités. Il y enseigna ensuite toute sa carrière, et fut élu à l'Académie des sciences de Saint-Pétersbourg [ORa].

Markov introduisit les chaînes qui portent son nom en 1906 [Mar06], dans le cadre d'une controverse avec le mathématicien moscovite Pavel Nekrassov. Celui-ci soutenait que la loi des grands nombres, selon laquelle une fréquence observée se rapproche d'une probabilité, exigeait que les expériences soient indépendantes, et il en tirait des conclusions sur le libre arbitre. Markov montra que la loi reste vraie pour des variables dépendantes, pourvu que la dépendance soit du type qu'il décrivait : chaque résultat ne dépend que du précédent [Hay13]. Ses chaînes étaient d'abord, pour lui, un contre-exemple théorique ; il ne les appliqua qu'une fois à des données réelles (note historique 4.2).
Markov était connu pour son indépendance d'esprit. En 1902, il protesta lorsque l'Académie, à la demande du tsar, annula l'élection de l'écrivain Maxime Gorki. En 1912, il demanda au Saint-Synode d'être excommunié, par solidarité avec Léon Tolstoï, qui l'avait été. En 1913, alors que l'Empire célébrait le tricentenaire de la dynastie des Romanov, il organisa plutôt la célébration d'un autre anniversaire : le bicentenaire de la loi des grands nombres, publiée par Jacques Bernoulli dans son *Ars conjectandi* en 1713 [Hay13]. Son fils, Andreï Andreïevitch Markov (1903–1979), fut lui aussi mathématicien, spécialiste de logique.

4.1 Rappels sur les probabilités

4.1.1 Probabilités conditionnelles

Une **expérience aléatoire** est une expérience dont on ne peut pas prédire le résultat avec certitude : lancer un dé, tirer une carte, observer le temps qu'il fera demain. L'ensemble des résultats possibles s'appelle l'**univers** de l'expérience, et un **événement** est un sous-ensemble de l'univers. On note $\mathbb{P}(A)$ la probabilité d'un événement A ; c'est un nombre de l'intervalle $[0, 1]$.

Il arrive souvent qu'une information partielle modifie la probabilité d'un événement. S'il pleut ce matin, la probabilité que l'autobus soit en retard n'est pas la même qu'un matin ensoleillé.

Définition 4.1 **Probabilité conditionnelle**

Soient A et B deux événements, avec $\mathbb{P}(B) > 0$. La **probabilité conditionnelle** de A sachant B est

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Autrement dit, $\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\,\mathbb{P}(B)$: pour que A et B se produisent tous les deux, il faut que B se produise, puis que A se produise sachant B. C'est la règle que l'on applique lorsqu'on multiplie les probabilités le long d'une branche d'un diagramme en arbre. La règle d'addition, qui consiste à additionner les probabilités des branches menant à un même résultat, correspond à la proposition suivante.

Proposition 4.1 **Formule des probabilités totales**

Soient $B_1, \dots, B_N$ des événements deux à deux incompatibles, de probabilités non nulles, dont la réunion est l'univers tout entier. Alors, pour tout événement A,

$$\mathbb{P}(A) = \mathbb{P}(A \mid B_1)\,\mathbb{P}(B_1) + \mathbb{P}(A \mid B_2)\,\mathbb{P}(B_2) + \cdots + \mathbb{P}(A \mid B_N)\,\mathbb{P}(B_N).$$

Démonstration. Les événements $A \cap B_1, \dots, A \cap B_N$ sont deux à deux incompatibles et leur réunion est A. Donc

$$\mathbb{P}(A) = \sum_{j=1}^{N} \mathbb{P}\big(A \cap B_j\big) = \sum_{j=1}^{N} \mathbb{P}\big(A \mid B_j\big)\,\mathbb{P}\big(B_j\big).$$

□

Exemple 4.1 **L'autobus du matin**

Un matin sur trois en moyenne, il pleut. Lorsqu'il pleut, l'autobus est en retard avec une probabilité de 0,4 ; sinon, avec une probabilité de 0,1. Notons P l'événement « il pleut » et R l'événement « l'autobus est en retard ». Le diagramme en arbre de la figure 4.2 et la formule des probabilités totales donnent

$$\mathbb{P}(R) = \mathbb{P}(R \mid P)\,\mathbb{P}(P) + \mathbb{P}\big(R \mid \overline{P}\big)\,\mathbb{P}\big(\overline{P}\big) = 0{,}4 \cdot \tfrac{1}{3} + 0{,}1 \cdot \tfrac{2}{3} = \tfrac{0{,}6}{3} = 0{,}2.$$

L'autobus est donc en retard un matin sur cinq.

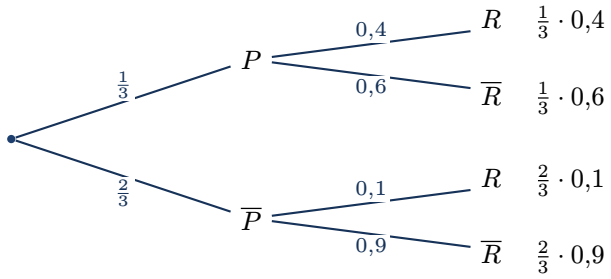


FIGURE 4.2 : Le diagramme en arbre de l'exemple 4.1. Les probabilités inscrites sur les branches du second niveau sont des probabilités conditionnelles ; on multiplie le long d'une branche et on additionne les branches qui mènent à R.

4.1.2 Variables aléatoires discrètes et espérance

Définition 4.2 **Variable aléatoire discrète**

Une **variable aléatoire discrète** X est une fonction qui associe un nombre à chaque résultat d'une expérience aléatoire, et qui ne prend qu'un nombre fini (ou une suite) de valeurs $x_1, x_2, \ldots$ La **loi** de X est la donnée des probabilités $p_i = \mathbb{P}(X = x_i)$; ces nombres sont positifs et leur somme vaut 1. L'**espérance** de X est la moyenne de ses valeurs, pondérée par leurs probabilités :

$$\mathbb{E}(X) = x_1 p_1 + x_2 p_2 + \cdots = \sum_i x_i\, p_i.$$

Si X est le résultat d'un dé équilibré, sa loi attribue la probabilité $\frac{1}{6}$ à chacune des valeurs $1, \ldots, 6$, et $\mathbb{E}(X) = \frac{1}{6}(1 + 2 + \cdots + 6) = 3{,}5$. L'espérance n'est pas nécessairement une valeur que X peut prendre : c'est la valeur moyenne que l'on observerait en répétant l'expérience un grand nombre de fois. Deux propriétés nous seront utiles. D'abord, l'espérance est *linéaire* : $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$, même lorsque X et Y ne sont pas indépendantes. Ensuite, si $\mathbb{1}_A$ est la variable **indicatrice** d'un événement A, qui vaut 1 si A se produit et 0 sinon, alors $\mathbb{E}(\mathbb{1}_A) = 1 \cdot \mathbb{P}(A) + 0 \cdot \mathbb{P}\big(\overline{A}\big) = \mathbb{P}(A)$. Ces deux propriétés permettent de compter *en moyenne* sans connaître la loi d'une somme : nous en verrons un exemple à la section 4.2.

Dans ce chapitre, les variables aléatoires représenteront l'*état* d'un système : le temps qu'il fait, le lieu de résidence d'une personne, la page sur laquelle se trouve un internaute. Nous numéroterons les états $1, 2, \ldots, N$, de sorte que la loi d'une telle variable est décrite par un vecteur de $\mathbb{R}^N$.

Définition 4.3 **Vecteur de probabilités**

Un vecteur $\mathbf{p} = (p_1, \ldots, p_N)$ de $\mathbb{R}^N$ est un **vecteur de probabilités** si ses composantes sont positives ou nulles et si leur somme vaut 1 :

$$p_i \geq 0 \quad \text{pour tout } i, \qquad p_1 + p_2 + \cdots + p_N = 1.$$

Si X est une variable aléatoire à valeurs dans $\{1, \ldots, N\}$, le vecteur $(\mathbb{P}(X = 1), \ldots, \mathbb{P}(X = N))$ est un vecteur de probabilités, que l'on appelle la loi de X.

Géométriquement, les vecteurs de probabilités de $\mathbb{R}^2$ forment le segment qui joint $(1, 0)$ à $(0, 1)$, et ceux de $\mathbb{R}^3$ forment le triangle plein de sommets $(1, 0, 0)$, $(0, 1, 0)$ et $(0, 0, 1)$. On appelle cet ensemble le **simplexe** de $\mathbb{R}^N$. En écrivant $\mathbf{e} = (1, 1, \ldots, 1)$, la condition sur la somme s'écrit aussi $\mathbf{e}^{\mathsf{T}}\mathbf{p} = 1$, une écriture qui nous sera utile pour les calculs matriciels.

Pour la classe 4.1 **Les probabilités conditionnelles au secondaire**

Les élèves dénombrent des résultats à l'aide de diagrammes en arbre dès le primaire. Au deuxième cycle du secondaire, ils recourent à l'espérance mathématique, mise en lien avec la moyenne pondérée, pour juger de l'équité d'un jeu, et ils prennent des décisions à partir de probabilités conditionnelles [Min16]. Le diagramme en arbre y joue le rôle que jouera ici la matrice de transition : les probabilités inscrites sur les branches du second niveau sont des probabilités conditionnelles, et la règle « multiplier le long d'une branche, additionner les branches » est la formule des probabilités totales. Un arbre devient toutefois rapidement illisible lorsque le

nombre d'étapes augmente : avec deux états, il compte déjà $2^{10} = 1024$ branches après dix étapes. C'est l'une des raisons d'introduire la notation matricielle.

4.2 Un premier modèle : la météo à deux états

Dans une ville imaginaire, on a observé pendant de nombreuses années que le temps qu'il fait un jour donné influence celui du lendemain :

— après un jour de beau temps, il fait beau le lendemain avec une probabilité de 0,8, et il pleut avec une probabilité de 0,2 ;

— après un jour de pluie, il fait beau le lendemain avec une probabilité de 0,4, et il pleut de nouveau avec une probabilité de 0,6.

Nous supposons que ces probabilités ne dépendent ni de la saison, ni du temps qu'il a fait les jours précédents : seul compte le temps de la veille. On résume la situation par un **diagramme de transition** (figure 4.3), dans lequel chaque état est un sommet et chaque flèche porte la probabilité de passer d'un état à l'autre.

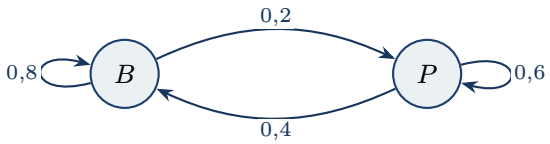


FIGURE 4.3 : Le diagramme de transition du modèle météorologique. L'état B représente un jour de beau temps et l'état P, un jour de pluie. Les probabilités des flèches qui *partent* d'un même état ont une somme égale à 1.

4.2.1 Prévoir le temps des jours suivants

Supposons qu'il fasse beau aujourd'hui, le jour 0. Notons X_n le temps qu'il fera dans n jours. Par hypothèse, $\mathbb{P}(X_1 = B) = 0{,}8$ et $\mathbb{P}(X_1 = P) = 0{,}2$. Pour le surlendemain, il faut distinguer selon le temps du lendemain ; la formule des probabilités totales donne

$$\begin{aligned}\mathbb{P}(X_2 = B) &= \mathbb{P}(X_2 = B \mid X_1 = B)\,\mathbb{P}(X_1 = B) + \mathbb{P}(X_2 = B \mid X_1 = P)\,\mathbb{P}(X_1 = P)\\ &= 0{,}8 \cdot 0{,}8 + 0{,}4 \cdot 0{,}2 = 0{,}72,\end{aligned}$$

et de même $\mathbb{P}(X_2 = P) = 0{,}2 \cdot 0{,}8 + 0{,}6 \cdot 0{,}2 = 0{,}28$. Le même calcul, répété, donne la probabilité de chaque état à l'étape $n+1$ à partir des probabilités à l'étape n :

$$\begin{cases}\mathbb{P}(X_{n+1} = B) = 0{,}8\,\mathbb{P}(X_n = B) + 0{,}4\,\mathbb{P}(X_n = P),\\ \mathbb{P}(X_{n+1} = P) = 0{,}2\,\mathbb{P}(X_n = B) + 0{,}6\,\mathbb{P}(X_n = P).\end{cases} \tag{4.1}$$

On reconnaît un produit matriciel. Notons $\mathbf{x}_n = \big(\mathbb{P}(X_n = B), \mathbb{P}(X_n = P)\big)$ la loi de X_n ; c'est un vecteur de probabilités. Le système (4.1) s'écrit

$$\mathbf{x}_{n+1} = M\mathbf{x}_n, \qquad \text{où} \qquad M = \begin{pmatrix} 0{,}8 & 0{,}4 \\ 0{,}2 & 0{,}6 \end{pmatrix}.$$

La matrice M s'appelle la **matrice de transition** du modèle. Sa première colonne contient les probabilités des transitions qui partent de l'état B, et sa seconde colonne, celles qui partent de l'état P ; le coefficient m_{ij} est la probabilité de passer *de l'état j à l'état i*. En appliquant la récurrence plusieurs fois, on obtient

$$\mathbf{x}_1 = M\mathbf{x}_0, \quad \mathbf{x}_2 = M\mathbf{x}_1 = M^2\mathbf{x}_0, \quad \ldots, \quad \mathbf{x}_n = M^n\mathbf{x}_0.$$

Prévoir le temps dans n jours revient donc à calculer une puissance de matrice : c'est ici que l'algèbre linéaire entre en jeu.

Remarque 4.1 **Lignes ou colonnes ?**

Nous plaçons les probabilités de transition qui partent d'un même état dans une *colonne*, de sorte que la loi évolue par $\mathbf{x}_{n+1} = M\mathbf{x}_n$, comme les vecteurs de l'annexe D. Beaucoup d'ouvrages de probabilités, ainsi que Wikipédia, utilisent la convention inverse : ils placent ces probabilités dans une *ligne*, écrivent la loi comme un vecteur ligne et la font évoluer par $_{n+1} = {}_nP$. Les deux conventions sont équivalentes, puisque $P = M^{\mathsf{T}}$, mais il faut repérer celle qu'utilise une source avant d'en reprendre les calculs. Dans notre convention, la somme de chaque *colonne* de M vaut 1.

Le tableau 4.1 donne les premières valeurs de $\mathbf{x}_n$ lorsqu'il fait beau le jour 0, puis lorsqu'il pleut le jour 0. Dans les deux cas, les probabilités se rapprochent rapidement de $\left(\frac{2}{3}, \frac{1}{3}\right) \approx (0{,}6667\,;\ 0{,}3333)$: après une semaine, le temps qu'il faisait le jour 0 n'a presque plus d'influence.

TABLE 4.1 : L'évolution de la loi $\mathbf{x}_n = M^n\mathbf{x}_0$ du modèle météorologique, selon le temps du jour 0. On a indiqué $\mathbb{P}(X_n = B)$; la probabilité de pluie s'obtient par complément.

n	0	1	2	3	4	5	6	7
beau le jour 0	1	0,8	0,72	0,688	0,6752	0,6701	0,6680	0,6672
pluie le jour 0	0	0,4	0,56	0,624	0,6496	0,6598	0,6639	0,6656

4.2.2 L'état d'équilibre

Si la loi $\mathbf{x}_n$ converge vers un vecteur $\mathbf{q}$, alors, en passant à la limite dans $\mathbf{x}_{n+1} = M\mathbf{x}_n$, on obtient $\mathbf{q} = M\mathbf{q}$. Le vecteur $\mathbf{q}$ ne change pas lorsqu'on lui applique la transition : c'est un **état d'équilibre**, aussi appelé **loi stationnaire**. En langage d'algèbre linéaire, $\mathbf{q}$ est un vecteur propre de M associé à la valeur propre 1 (définition D.10) ; en langage du chapitre 2, c'est un point fixe de l'application $\mathbf{x} \mapsto M\mathbf{x}$.

Cherchons-le. L'équation $(M - I)\mathbf{q} = \mathbf{0}$ s'écrit

$$\begin{cases} -0{,}2\,q_1 + 0{,}4\,q_2 = 0 \\ 0{,}2\,q_1 - 0{,}4\,q_2 = 0, \end{cases}$$

deux équations équivalentes, qui donnent $q_1 = 2q_2$. La condition $q_1 + q_2 = 1$ fournit ensuite l'unique vecteur de probabilités $\mathbf{q} = \left(\frac{2}{3}, \frac{1}{3}\right)$. À long terme, il fait beau deux jours sur trois.

Pourquoi la convergence est-elle si rapide ? Le polynôme caractéristique de M est $\lambda^2 - 1{,}4\,\lambda + 0{,}4 = (\lambda - 1)(\lambda - 0{,}4)$; la seconde valeur propre est $\lambda_2 = 0{,}4$, associée au vecteur propre $\mathbf{v}_2 = (1, -1)$. Comme à l'exemple D.8, écrivons le vecteur de départ dans la base propre $\{\mathbf{q}, \mathbf{v}_2\}$. Pour $\mathbf{x}_0 = (1, 0)$, on a $\mathbf{x}_0 = \mathbf{q} + \frac{1}{3}\mathbf{v}_2$, d'où

$$\mathbf{x}_n = M^n\mathbf{x}_0 = M^n\mathbf{q} + \tfrac{1}{3}M^n\mathbf{v}_2 = \mathbf{q} + \tfrac{1}{3}\,(0{,}4)^n\,\mathbf{v}_2.$$

L'écart à l'équilibre est multiplié par 0,4 chaque jour : il est d'environ 0,0005 après une semaine, ce que confirme le tableau 4.1. La seconde valeur propre gouverne donc la vitesse de convergence ; nous retrouverons ce fait à la section 4.6.

Exemple 4.2 **Combien de beaux jours dans la semaine ?**

Il fait beau aujourd'hui. Combien de jours de beau temps peut-on espérer pendant les sept jours suivants ? Notons Y ce nombre. On peut écrire $Y = \mathbb{1}_1 + \mathbb{1}_2 + \cdots + \mathbb{1}_7$, où $\mathbb{1}_n$ est l'indicatrice de l'événement « il fait beau le jour n ». Ces variables ne sont pas indépendantes, mais la linéarité de l'espérance suffit :

$$\mathbb{E}(Y) = \sum_{n=1}^{7} \mathbb{E}(\mathbb{1}_n) = \sum_{n=1}^{7} \mathbb{P}(X_n = B) \approx 0{,}8 + 0{,}72 + 0{,}688 + \cdots + 0{,}6672 \approx 4{,}89.$$

On peut espérer près de cinq beaux jours, un peu plus que les $7 \cdot \frac{2}{3} \approx 4{,}67$ jours qu'on prévoirait à l'équilibre, puisque la semaine commence par un beau jour.

4.2.3 Le calcul avec Python

Le programme suivant calcule quelques puissances de M, puis l'état d'équilibre à l'aide de `np.linalg.eig`. La fonction renvoie des vecteurs propres unitaires (de norme 1) ; il faut donc diviser le vecteur propre associé à 1 par la somme de ses composantes pour obtenir un vecteur de probabilités (exercice 4.9).

Programme 4.1 : Le modèle météorologique : puissances de M et état d'équilibre.

```
import numpy as np
np.set_printoptions(precision=4, suppress=True)

# colonnes : départ d'un beau jour, puis d'un jour de pluie
M = np.array([[0.8, 0.4],
              [0.2, 0.6]])
x0 = np.array([1.0, 0.0])     # aujourd'hui, il fait beau

for n in [1, 2, 5, 10, 20]:
    print(n, np.linalg.matrix_power(M, n) @ x0)

valeurs, vecteurs = np.linalg.eig(M)
print("valeurs propres :", valeurs)
v = vecteurs[:, np.argmax(valeurs)]   # associé à 1
print("équilibre :", v / v.sum())     # normalisé : somme = 1
```

Programme 4.2 : Sortie.

```
1 [0.8 0.2]
2 [0.72 0.28]
5 [0.6701 0.3299]
10 [0.6667 0.3333]
20 [0.6667 0.3333]
valeurs propres : [1.  0.4]
équilibre : [0.6667 0.3333]
```

L'équilibre a aussi une interprétation en termes de *fréquences*. Si l'on simule le temps sur un grand nombre de jours, en tirant chaque jour le temps du lendemain selon les probabilités de transition, la proportion des jours de beau temps devrait se rapprocher de $\frac{2}{3}$. C'est ce que vérifie le programme suivant, qui simule dix mille jours.

Programme 4.3 : Simulation du modèle météorologique sur dix mille jours.

```python
import numpy as np
rng = np.random.default_rng(2026)      # graine fixée

M = np.array([[0.8, 0.4],
              [0.2, 0.6]])
etat, beaux = 0, 0                     # 0 : beau ; 1 : pluie
n = 10_000
for jour in range(n):
    # tirage selon la colonne de l'état actuel
    etat = rng.choice(2, p=M[:, etat])
    beaux += (etat == 0)
print("proportion de jours de beau temps :", beaux / n)
```

Programme 4.4 : Sortie.

```
proportion de jours de beau temps : 0.6644
```

La proportion observée, 0,6644, est proche de $\frac{2}{3}$; elle varierait légèrement avec une autre graine. Cette double interprétation de l'équilibre, comme *limite des probabilités* et comme *proportion du temps passé dans chaque état*, est au cœur de l'algorithme PageRank.

Outil 4.1 **Puissances de matrices sans programmer**

Pour une petite matrice, il n'est pas nécessaire d'écrire un programme. Dans WolframAlpha, la requête `{{0.8, 0.4}, {0.2, 0.6}}^10` calcule M^{10}, et `eigenvectors {{0.8, 0.4}, {0.2, 0.6}}` donne les vecteurs propres. Dans GeoGebra, on définit `M = {{0.8, 0.4}, {0.2, 0.6}}` et l'on calcule `M^n`, avec un curseur n : en faisant varier n, on voit les deux colonnes de M^n devenir presque égales à $\left(\frac{2}{3}, \frac{1}{3}\right)$. Ce dernier constat est instructif : si toutes les colonnes de M^n sont proches de $\mathbf{q}$, alors $M^n\mathbf{x}_0$ est proche de $\mathbf{q}$ pour *tout* vecteur de probabilités $\mathbf{x}_0$, ce qui explique que le point de départ soit oublié. Un tableur convient aussi : chaque ligne contient la loi d'un jour, calculée à partir de la ligne précédente par les deux formules du système (4.1).

4.3 Les chaînes de Markov

4.3.1 Définition

Le modèle météorologique repose sur une hypothèse précise : pour prévoir le temps de demain, il suffit de connaître le temps d'aujourd'hui ; connaître aussi celui des jours précédents n'apporte aucune information supplémentaire. C'est cette absence de mémoire, au-delà de l'étape présente, qui définit une chaîne de Markov.

Définition 4.4 **Chaîne de Markov**

Une suite de variables aléatoires $X_0, X_1, X_2, \ldots$ à valeurs dans un ensemble fini d'**états** $\{1, 2, \ldots, N\}$ est une **chaîne de Markov** si, pour tout n et pour tous les états $i, j, i_0, \ldots, i_{n-1}$,

$$\mathbb{P}(X_{n+1} = i \mid X_n = j,\ X_{n-1} = i_{n-1},\ \ldots,\ X_0 = i_0) = \mathbb{P}(X_{n+1} = i \mid X_n = j).$$

Cette égalité s'appelle la **propriété de Markov**. La chaîne est **homogène** si cette probabilité ne dépend pas de n ; on la note alors m_{ij}, et la matrice $M = (m_{ij})$ de taille $N \times N$ s'appelle la **matrice de transition** de la chaîne.

Toutes les chaînes de ce chapitre sont homogènes. La propriété de Markov n'interdit pas le hasard de dépendre du passé : elle exige seulement que toute l'information utile sur le passé soit contenue dans l'état présent. Une suite de lancers d'une pièce de monnaie est une chaîne de Markov d'un type particulier, puisque chaque lancer ne dépend même pas du précédent : toutes les colonnes de sa matrice de transition sont égales à $(\frac{1}{2}, \frac{1}{2})$. À l'opposé, nous verrons à la section 4.3.3 un modèle dont l'avenir dépend des deux derniers états.

Proposition 4.2 **Évolution de la loi**

Soit (X_n) une chaîne de Markov homogène de matrice de transition M, et soit $\mathbf{x}_n$ la loi de X_n. Alors $\mathbf{x}_{n+1} = M\mathbf{x}_n$ pour tout n, et par conséquent $\mathbf{x}_n = M^n\mathbf{x}_0$.

Démonstration. Les événements $\{X_n = 1\}, \ldots, \{X_n = N\}$ sont deux à deux incompatibles et leur réunion est l'univers. Pour un état i, la formule des probabilités totales (proposition 4.1) donne

$$\mathbb{P}(X_{n+1} = i) = \sum_{j=1}^{N} \mathbb{P}(X_{n+1} = i \mid X_n = j)\,\mathbb{P}(X_n = j) = \sum_{j=1}^{N} m_{ij}\,\mathbb{P}(X_n = j),$$

où l'on omet les termes pour lesquels $\mathbb{P}(X_n = j) = 0$. Le membre de droite est la i e composante du produit $M\mathbf{x}_n$ (équation (D.2)). La seconde affirmation s'obtient par récurrence sur n. □

Les coefficients de la matrice M^n ont eux aussi une interprétation : en appliquant la proposition à $\mathbf{x}_0 = \mathbf{e}_j$, le vecteur dont seule la j e composante est non nulle et vaut 1, on voit que la j e colonne de M^n est la loi de X_n lorsque la chaîne part de l'état j. Le coefficient $(M^n)_{ij}$ est donc la probabilité de passer de l'état j à l'état i en n étapes.

Pour la classe 4.2 **Graphes et matrices au secondaire**

Le diagramme de transition d'une chaîne de Markov est un *graphe orienté valué* : ses sommets sont les états, ses arêtes sont orientées et portent un poids. La progression des apprentissages prévoit une introduction à la théorie des graphes au deuxième cycle du secondaire, où les élèves construisent des graphes orientés et valués pour représenter une situation [Min16]. Dans la séquence Technico-sciences, ils utilisent aussi des matrices comme registre de représentation, et effectuent des multiplications matricielles, notamment pour traiter des données à la manière d'un tableur. Une chaîne de Markov réunit ces deux registres : le graphe pour décrire la situation, la matrice pour calculer. Passer de l'un à l'autre (exercice 4.14) est une

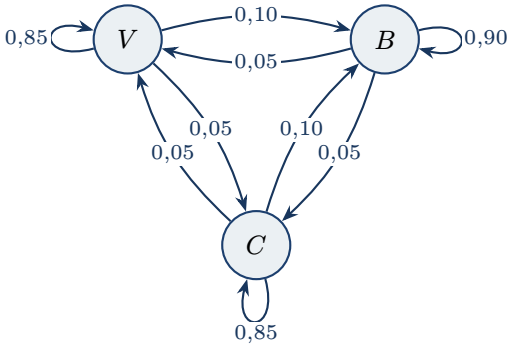
tâche accessible, qui oblige à préciser ce que représente chaque ligne et chaque colonne.

4.3.2 Ville, banlieue et campagne

Considérons maintenant une population répartie entre trois milieux de vie : la ville (V), la banlieue (B) et la campagne (C). Chaque année, on observe les déménagements suivants :

- parmi les citadins, 85 % restent en ville, 10 % s'installent en banlieue et 5 % à la campagne ;
- parmi les banlieusards, 90 % restent en banlieue, 5 % partent pour la ville et 5 % pour la campagne ;
- parmi les ruraux, 85 % restent à la campagne, 10 % s'installent en banlieue et 5 % en ville.

Nous supposons que la population totale est constante et que ces taux ne changent pas d'une année à l'autre. Le lieu de résidence d'une personne choisie au hasard est alors une chaîne de Markov à trois états, dont le diagramme et la matrice de transition sont donnés à la figure 4.4. Les proportions de la population dans chaque milieu forment un vecteur de probabilités $\mathbf{x}_n$, qui évolue selon $\mathbf{x}_{n+1} = M\mathbf{x}_n$.



$$M = \begin{pmatrix} 0{,}85 & 0{,}05 & 0{,}05 \\ 0{,}10 & 0{,}90 & 0{,}10 \\ 0{,}05 & 0{,}05 & 0{,}85 \end{pmatrix}$$

colonnes : départ de V, de B, de C
lignes : arrivée en V, en B, en C

FIGURE 4.4 : Le modèle de migration entre la ville (V), la banlieue (B) et la campagne (C). Chaque colonne de M correspond aux flèches qui partent d'un même état.

Si la population se répartit aujourd'hui en 50 % en ville, 30 % en banlieue et 20 % à la campagne, alors $\mathbf{x}_0 = (0{,}5\,;\ 0{,}3\,;\ 0{,}2)$ et

$$\mathbf{x}_1 = M\mathbf{x}_0 = \begin{pmatrix} 0{,}85 \cdot 0{,}5 + 0{,}05 \cdot 0{,}3 + 0{,}05 \cdot 0{,}2 \\ 0{,}10 \cdot 0{,}5 + 0{,}90 \cdot 0{,}3 + 0{,}10 \cdot 0{,}2 \\ 0{,}05 \cdot 0{,}5 + 0{,}05 \cdot 0{,}3 + 0{,}85 \cdot 0{,}2 \end{pmatrix} = \begin{pmatrix} 0{,}45 \\ 0{,}34 \\ 0{,}21 \end{pmatrix}.$$

Le tableau 4.2 montre l'évolution sur un plus long horizon : la banlieue se peuple au détriment de la ville, et les proportions se stabilisent autour de $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$.

TABLE 4.2 : L'évolution de la répartition $\mathbf{x}_n = M^n\mathbf{x}_0$ de la population dans le modèle de migration.

n (années)	0	1	2	5	10	20	50
ville	0,50	0,45	0,41	0,3319	0,2768	0,2529	0,2500
banlieue	0,30	0,34	0,372	0,4345	0,4785	0,4977	0,5000
campagne	0,20	0,21	0,218	0,2336	0,2446	0,2494	0,2500

L'état d'équilibre est la solution de $(M - I)\mathbf{q} = \mathbf{0}$ dont les composantes ont une somme

égale à 1. La méthode de Gauss–Jordan (section D.5) appliquée à

$$M - I = \begin{pmatrix} -0{,}15 & 0{,}05 & 0{,}05 \\ 0{,}10 & -0{,}10 & 0{,}10 \\ 0{,}05 & 0{,}05 & -0{,}15 \end{pmatrix}$$

montre que son noyau est la droite engendrée par $(1, 2, 1)$; en divisant par la somme des composantes, on obtient $\mathbf{q} = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$. À long terme, la moitié de la population vit en banlieue.

Exemple 4.3 **Une formule exacte pour $\mathbf{x}_n$**

La matrice de ce modèle a une structure particulière. Notons Q la matrice 3×3 dont les trois colonnes sont égales à $\mathbf{q}$, c'est-à-dire $Q = \mathbf{q}\,\mathbf{e}^{\mathsf{T}}$. On vérifie que

$$M = 0{,}8\,I + 0{,}2\,Q\,:$$

par exemple, la première colonne de $0{,}8\,I + 0{,}2\,Q$ vaut $(0{,}8 + 0{,}05\,;\ 0{,}1\,;\ 0{,}05)$. Autrement dit, chaque année, une personne garde son lieu de résidence avec une probabilité de $0{,}8$; avec une probabilité de $0{,}2$, elle « tire au sort » un nouveau milieu de vie selon la loi $\mathbf{q}$, ce qui peut la ramener au même endroit. Soit maintenant $\mathbf{v}$ un vecteur dont la somme des composantes est nulle. Alors $\mathbf{e}^{\mathsf{T}}\mathbf{v} = 0$, donc $Q\mathbf{v} = \mathbf{q}\,(\mathbf{e}^{\mathsf{T}}\mathbf{v}) = \mathbf{0}$ et $M\mathbf{v} = 0{,}8\,\mathbf{v}$. Tous les vecteurs du plan d'équation $v_1 + v_2 + v_3 = 0$ sont donc des vecteurs propres associés à la valeur propre $0{,}8$. Comme $\mathbf{x}_0 - \mathbf{q}$ est dans ce plan, on obtient

$$\mathbf{x}_n - \mathbf{q} = M^n(\mathbf{x}_0 - \mathbf{q}) = (0{,}8)^n(\mathbf{x}_0 - \mathbf{q}), \qquad \text{soit} \qquad \mathbf{x}_n = \mathbf{q} + (0{,}8)^n(\mathbf{x}_0 - \mathbf{q}).$$

L'écart à l'équilibre diminue de $20\,\%$ par année : il faut une dizaine d'années pour le réduire à environ le dixième de sa valeur initiale, puisque $(0{,}8)^{10} \approx 0{,}107$. Nous retrouverons cette idée de « tirage au sort » dans la matrice de Google, à la section 4.5.3.

4.3.3 Et si l'avenir dépendait de plusieurs jours ?

Revenons à la météo. Des observations plus fines pourraient montrer que le temps de demain dépend des *deux* derniers jours : après deux jours de beau temps, le beau temps est bien installé, alors qu'après un seul beau jour suivant un jour de pluie, il l'est moins. Supposons par exemple que la probabilité qu'il fasse beau demain soit donnée par le tableau suivant.

hier	aujourd'hui	$\mathbb{P}$(beau demain)
B	B	0,9
P	B	0,7
B	P	0,5
P	P	0,3

La suite (X_n) des temps observés n'est plus une chaîne de Markov : la probabilité qu'il fasse beau demain sachant qu'il fait beau aujourd'hui n'est pas déterminée, puisqu'elle vaut 0,9 ou 0,7 selon le temps de la veille.

On récupère cependant une chaîne de Markov en changeant d'états. Considérons le *couple* $Y_n = (X_{n-1}, X_n)$ formé du temps d'hier et de celui d'aujourd'hui. Il y a quatre états, BB, PB, BP et PP, et la connaissance de Y_n suffit pour prévoir Y_{n+1} : depuis l'état

(h, a), on ne peut passer qu'à (a, B) ou à (a, P). Par exemple, depuis PB (pluie hier, beau aujourd'hui), on passe à BB avec une probabilité de 0,7, et à BP avec une probabilité de 0,3. La matrice de transition, dans l'ordre BB, PB, BP, PP, est

$$M_2 = \begin{pmatrix} 0{,}9 & 0{,}7 & 0 & 0 \\ 0 & 0 & 0{,}5 & 0{,}3 \\ 0{,}1 & 0{,}3 & 0 & 0 \\ 0 & 0 & 0{,}5 & 0{,}7 \end{pmatrix}.$$

Elle est plus grande, et elle contient beaucoup de zéros, parce que certaines transitions sont impossibles : depuis BB, on ne peut pas passer à PP en un jour. Son état d'équilibre est $\mathbf{q} = \left(\frac{21}{32}, \frac{3}{32}, \frac{3}{32}, \frac{5}{32}\right)$ (exercice 4.10). La probabilité qu'il fasse beau un jour donné, à long terme, est donc $q_{BB} + q_{PB} = \frac{24}{32} = \frac{3}{4}$.

Cette construction est générale : un phénomène dont l'avenir dépend des k derniers états se ramène à une chaîne de Markov dont les états sont des suites de k états. Le prix à payer est la taille de la matrice : avec N états de départ, on obtient N^k nouveaux états. Les premiers générateurs automatiques de texte fonctionnaient ainsi : l'état est formé des k derniers mots, et le mot suivant est tiré au hasard selon des fréquences observées dans un grand corpus. Avec $k = 1$ ou 2, le texte produit est localement plausible, mais il perd le fil après quelques mots ; augmenter k améliore la cohérence, mais le nombre d'états devient vite énorme.

Note historique 4.2 Les voyelles d'Eugène Onéguine

En 1913, Markov (note historique 4.1) a appliqué son modèle aux 20 000 premières lettres du roman en vers *Eugène Onéguine* d'Alexandre Pouchkine, en ne retenant que deux états : voyelle ou consonne [Mar13]. Il compta 8638 voyelles ; une voyelle était suivie d'une voyelle 1104 fois, soit avec une fréquence de 0,128, alors qu'une consonne était suivie d'une voyelle avec une fréquence de 0,663. La chaîne correspondante a pour matrice de transition $\begin{pmatrix} 0{,}128 & 0{,}663 \\ 0{,}872 & 0{,}337 \end{pmatrix}$, et son état d'équilibre attribue aux voyelles la probabilité $0{,}663/(0{,}872 + 0{,}663) \approx 0{,}432$, soit exactement la proportion $8638/20\,000$ observée dans le texte. Ce calcul, fait à la main, est la première application d'une chaîne de Markov à des données réelles.

4.4 Matrices stochastiques et états d'équilibre

Dans les deux exemples précédents, la loi $\mathbf{x}_n$ converge vers un unique état d'équilibre, quel que soit l'état de départ. Est-ce toujours le cas ? Pour répondre, nous dégageons d'abord les propriétés communes à toutes les matrices de transition.

Définition 4.5 Matrice stochastique

Une matrice carrée M est **stochastique** si ses coefficients sont positifs ou nuls et si la somme des coefficients de chacune de ses colonnes vaut 1. Autrement dit, chaque colonne de M est un vecteur de probabilités, ce qui s'écrit aussi $\mathbf{e}^{\mathsf{T}} M = \mathbf{e}^{\mathsf{T}}$.

La matrice de transition d'une chaîne de Markov est stochastique : sa j^{e} colonne est la loi de l'état suivant lorsqu'on part de l'état j. Réciproquement, toute matrice stochastique est la matrice de transition d'une chaîne de Markov.

Proposition 4.3 **Une matrice stochastique conserve les lois**

Si M est stochastique et si $\mathbf{v}$ est un vecteur de probabilités, alors $M\mathbf{v}$ est un vecteur de probabilités. Par conséquent, $M^n\mathbf{v}$ est un vecteur de probabilités pour tout $n \in \mathbb{N}$.

La démonstration fait l'objet de l'exercice 4.7 ; on peut la lire dans la notation matricielle : $\mathbf{e}^\intercal(M\mathbf{v}) = (\mathbf{e}^\intercal M)\mathbf{v} = \mathbf{e}^\intercal\mathbf{v} = 1$. La proposition confirme que le modèle est cohérent : $\mathbf{x}_n = M^n\mathbf{x}_0$ est toujours une loi.

4.4.1 La valeur propre 1

Théorème 4.4 **Valeur propre d'une matrice stochastique**

Si M est une matrice stochastique, alors 1 est une valeur propre de M.

Démonstration. Les matrices M et $M^\intercal$ ont le même polynôme caractéristique, puisque $\det(M^\intercal - \lambda I) = \det\left((M - \lambda I)^\intercal\right) = \det(M - \lambda I)$ (proposition D.1) ; elles ont donc les mêmes valeurs propres. Or, en transposant l'égalité $\mathbf{e}^\intercal M = \mathbf{e}^\intercal$, on obtient $M^\intercal\mathbf{e} = \mathbf{e}$: la i e composante de $M^\intercal\mathbf{e}$ est la somme des coefficients de la i e colonne de M, qui vaut 1. Le vecteur $\mathbf{e}$ est donc un vecteur propre de $M^\intercal$ associé à la valeur propre 1, et 1 est aussi une valeur propre de M. □

Ce théorème garantit l'existence d'un vecteur non nul $\mathbf{v}$ tel que $M\mathbf{v} = \mathbf{v}$, mais pas que ce vecteur ait des composantes positives : un vecteur propre peut toujours être multiplié par -1, et il pourrait aussi avoir des composantes de signes différents. Le résultat suivant, que nous admettrons, règle cette question. C'est une conséquence du théorème de Perron–Frobenius, qui porte sur les matrices à coefficients positifs [LM06].

Théorème 4.5 **Perron–Frobenius, cas des matrices stochastiques**

Soit M une matrice stochastique.

(i) Toute valeur propre λ de M, réelle ou complexe, vérifie $|\lambda| \leq 1$.

(ii) Il existe un vecteur de probabilités $\mathbf{q}$ tel que $M\mathbf{q} = \mathbf{q}$.

Toute chaîne de Markov à un nombre fini d'états possède donc au moins un état d'équilibre. La première affirmation est d'ailleurs facile à démontrer pour les valeurs propres réelles (exercice 4.8).

4.4.2 Deux chaînes qui ne convergent pas

L'existence d'un équilibre ne suffit pas à garantir que la loi $\mathbf{x}_n$ s'en rapproche. Deux exemples très simples le montrent.

Exemple 4.4 **Une chaîne qui ne bouge pas**

La matrice identité I de taille 2 est stochastique : chaque état est conservé avec probabilité 1. Tout vecteur de probabilités est un état d'équilibre, et $\mathbf{x}_n = \mathbf{x}_0$ pour tout n. L'équilibre n'est pas unique, et la limite dépend du point de départ. Ce comportement se produit chaque fois que la chaîne contient deux groupes d'états qui ne communiquent pas entre eux.

Exemple 4.5 **Une chaîne qui oscille**

La matrice $S = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ est stochastique : l'état change à chaque étape. Son unique état d'équilibre est $(\frac{1}{2}, \frac{1}{2})$, mais si l'on part de l'état 1, la loi alterne entre $(1, 0)$ et $(0, 1)$ et ne converge pas. La cause apparaît dans les valeurs propres : S a pour valeurs propres 1 et -1, et la composante de $\mathbf{x}_0$ dans la direction propre associée à -1 change de signe à chaque étape au lieu de disparaître. Une chaîne de ce type est dite **périodique**. Remarquons toutefois que la *proportion du temps* passé dans chaque état converge, elle, vers $\frac{1}{2}$.

Dans les deux cas, la valeur propre 1 n'est pas la seule valeur propre de module 1. Il suffit d'exclure ces situations pour obtenir la convergence. Une hypothèse commode est la suivante.

Définition 4.6 **Matrice stochastique régulière**

Une matrice stochastique M est **régulière** s'il existe un entier $k \geq 1$ tel que tous les coefficients de M^k soient strictement positifs.

Puisque $(M^k)_{ij}$ est la probabilité de passer de j à i en k étapes, M est régulière lorsqu'il existe un nombre d'étapes après lequel tous les états sont accessibles depuis tous les états. Les matrices du modèle météorologique et du modèle de migration sont régulières dès $k = 1$. Les matrices I et S ne le sont pas : toutes les puissances de I sont égales à I, et celles de S valent I ou S.

4.4.3 Le théorème de convergence

Théorème 4.6 **Convergence vers l'équilibre**

Soit M une matrice stochastique régulière de taille N. Alors :

(i) 1 est une valeur propre simple de M, et toute autre valeur propre λ vérifie $|\lambda| < 1$;

(ii) il existe un unique vecteur de probabilités $\mathbf{q}$ tel que $M\mathbf{q} = \mathbf{q}$, et ses composantes sont strictement positives ;

(iii) pour tout vecteur de probabilités $\mathbf{x}_0$, la suite $\mathbf{x}_n = M^n\mathbf{x}_0$ converge vers $\mathbf{q}$.

Les affirmations (i) et (ii) sont, elles aussi, des conséquences du théorème de Perron–Frobenius [LM06], que nous admettrons. Nous démontrons (iii) sous l'hypothèse supplémentaire que M est diagonalisable, c'est-à-dire qu'il existe une base de $\mathbb{R}^N$ formée de vecteurs propres de M (théorème D.3) ; la démonstration générale suit la même idée, mais demande des outils d'algèbre linéaire qui dépassent le cadre de ce livre.

Démonstration. [Démonstration de (iii) lorsque M est diagonalisable] Soit $\{\mathbf{q}, \mathbf{v}_2, \dots, \mathbf{v}_N\}$ une base de vecteurs propres, associés aux valeurs propres $1, \lambda_2, \dots, \lambda_N$, avec $|\lambda_i| < 1$ pour $i \geq 2$ par (i). Écrivons le vecteur de départ dans cette base :

$$\mathbf{x}_0 = \alpha_1\mathbf{q} + \alpha_2\mathbf{v}_2 + \cdots + \alpha_N\mathbf{v}_N.$$

Puisque $M^n\mathbf{v} = \lambda^n\mathbf{v}$ pour tout vecteur propre $\mathbf{v}$ associé à λ (exercice 4.4),

$$\mathbf{x}_n = M^n\mathbf{x}_0 = \alpha_1\mathbf{q} + \alpha_2\lambda_2^n\mathbf{v}_2 + \cdots + \alpha_N\lambda_N^n\mathbf{v}_N.$$

Chaque λ_i^n tend vers 0 lorsque $n \to \infty$, donc $\mathbf{x}_n \to \alpha_1 \mathbf{q}$. Il reste à voir que $\alpha_1 = 1$. Par la proposition 4.3, la somme des composantes de $\mathbf{x}_n$ vaut 1 pour tout n ; à la limite, celle de $\alpha_1 \mathbf{q}$ vaut aussi 1. Comme celle de $\mathbf{q}$ vaut 1, on obtient $\alpha_1 = 1$. □

La démonstration montre aussi à quelle *vitesse* la convergence se produit : l'écart $\mathbf{x}_n - \mathbf{q}$ est une combinaison de termes en λ_i^n, et c'est le plus grand des modules $|\lambda_2|, \dots, |\lambda_N|$ qui décide de la rapidité avec laquelle il disparaît. Pour le modèle météorologique, ce module vaut 0,4 ; pour le modèle de migration, il vaut 0,8, et la convergence est plus lente, comme le montrent les tableaux 4.1 et 4.2.

Remarque 4.2 **Proportion du temps et temps de retour**

L'état d'équilibre a deux interprétations, que la simulation de la section 4.2 illustrait. La première, que donne le théorème 4.6, est la limite des probabilités : $\mathbb{P}(X_n = i) \to q_i$. La seconde concerne une seule trajectoire de la chaîne : la proportion des étapes $0, 1, \dots, n-1$ passées dans l'état i tend vers q_i lorsque $n \to \infty$, avec probabilité 1. Ce résultat, que nous admettrons, est le **théorème ergodique** des chaînes de Markov, une version de la loi des grands nombres pour des étapes dépendantes ; il reste vrai pour les chaînes périodiques, comme celle de l'exemple 4.5, pourvu que l'équilibre soit unique. On en déduit une troisième interprétation : si la chaîne passe en moyenne une fraction q_i du temps dans l'état i, elle y revient en moyenne toutes les $1/q_i$ étapes. Dans le modèle météorologique, un jour de pluie revient en moyenne tous les 3 jours, et un jour de beau temps tous les 1,5 jour.

Pour la classe 4.3 **Simuler une chaîne de Markov en classe**

La comparaison entre probabilités fréquentielles et théoriques est au cœur de l'enseignement des probabilités dès le premier cycle [Min16]. Une chaîne de Markov s'y prête bien. Pour le modèle météorologique, on peut utiliser un dé à dix faces, ou une liste de nombres aléatoires : s'il fait beau, on tire « pluie » pour les résultats 1 et 2 ; s'il pleut, on tire « beau » pour les résultats 1 à 4. Chaque équipe simule un mois, puis on met en commun les résultats de la classe. La proportion de jours de beau temps se rapproche de $\frac{2}{3}$, et la discussion peut porter sur une question qui n'a rien d'évident : pourquoi la proportion ne dépend-elle pas du temps qu'il faisait le premier jour ? Les élèves peuvent aussi constater que les jours de pluie arrivent en séquences, contrairement à ce que donnerait une suite de tirages indépendants ; c'est une occasion de discuter de la conception erronée selon laquelle le hasard produirait des résultats « bien mélangés ».

4.5 L'algorithme PageRank

Un moteur de recherche classe les pages qui répondent à une requête selon deux types de critères. Les premiers portent sur le *contenu* de la page : la présence des mots cherchés, leur position dans le titre ou dans le texte, leur proximité. Les seconds portent sur la *popularité* de la page, mesurée à partir de la structure des liens du Web. L'algorithme PageRank, présenté par Sergey Brin et Larry Page en 1998 [BP98 ; Pag+99], relève de ce second type. Nous suivons ici la présentation de [Ger08].

4.5.1 L'internaute aléatoire

Imaginons un internaute qui navigue sans but précis. Lorsqu'il se trouve sur une page qui contient k liens vers d'autres pages, il clique sur l'un d'eux au hasard, chacun avec la probabilité $1/k$. Si cette promenade dure assez longtemps, la proportion du temps qu'il passe sur une page mesure l'importance de cette page : une page vers laquelle pointent de nombreux liens, ou des liens issus de pages elles-mêmes très visitées, sera souvent visitée.

Considérons le mini-web de la figure 4.1, et supposons que l'internaute parte de la page 3. La page 3 contient trois liens, vers les pages 2, 4 et 5 ; après un clic, il se trouve donc sur chacune de ces pages avec la probabilité $\frac{1}{3}$, et sur les pages 1 et 3 avec la probabilité 0. Pour se trouver sur la page 1 après deux clics, il doit être passé par la page 2 ou par la page 5, qui contiennent chacune deux liens dont l'un mène à la page 1. La formule des probabilités totales donne

$$\mathbb{P}(X_2 = 1) = \tfrac{1}{2} \cdot \mathbb{P}(X_1 = 2) + \tfrac{1}{2} \cdot \mathbb{P}(X_1 = 5) = \tfrac{1}{2} \cdot \tfrac{1}{3} + \tfrac{1}{2} \cdot \tfrac{1}{3} = \tfrac{1}{3}.$$

On reconnaît la démarche de la section 4.2 : la page sur laquelle se trouve l'internaute après n clics est une chaîne de Markov, dont les états sont les pages. Sa matrice de transition A a pour coefficient $a_{ij} = 1/k_j$ si la page j, qui contient k_j liens, pointe vers la page i, et $a_{ij} = 0$ sinon :

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

Chaque colonne de A décrit les liens qui *sortent* d'une page, et chaque ligne, les liens qui *entrent* dans une page. Avec $\mathbf{p}_0 = (0, 0, 1, 0, 0)$, on retrouve $\mathbf{p}_1 = A\mathbf{p}_0 = (0, \frac{1}{3}, 0, \frac{1}{3}, \frac{1}{3})$ et $\mathbf{p}_2 = A\mathbf{p}_1 = (\frac{1}{3}, \frac{1}{6}, \frac{1}{3}, 0, \frac{1}{6})$. Le programme suivant construit la matrice A à partir de la liste des liens, calcule quelques lois $\mathbf{p}_n = A^n\mathbf{p}_0$, puis l'état d'équilibre.

Programme 4.5 : Le PageRank du mini-web de la figure 4.1.

```
import numpy as np
np.set_printoptions(precision=4, suppress=True)

def matrice_transition(liens, N):
    """Matrice de transition d'un mini-web de N pages.
    liens[j] : pages vers lesquelles pointe la page j."""
    A = np.zeros((N, N))
    for j, sorties in liens.items():
        for i in sorties:
            # les pages sont numérotées à partir de 1
            A[i - 1, j - 1] = 1 / len(sorties)
    return A

liens = {1: [2], 2: [1, 3], 3: [2, 4, 5], 4: [3, 5], 5: [1, 2]}
A = matrice_transition(liens, 5)
p = np.array([0, 0, 1, 0, 0])          # départ : page 3
for n in [1, 2, 10, 50]:
    print(n, np.linalg.matrix_power(A, n) @ p)

valeurs, vecteurs = np.linalg.eig(A)
k = np.argmin(abs(valeurs - 1))        # valeur propre 1
q = np.real(vecteurs[:, k])
q = q / q.sum()
```

```
print("PageRank :", q)
print("fois 55  :", 55 * q)
```

Programme 4.6 : Sortie.

```
1 [0.     0.3333 0.     0.3333 0.3333]
2 [0.3333 0.1667 0.3333 0.     0.1667]
10 [0.2673 0.3166 0.254  0.0587 0.1033]
50 [0.2364 0.3636 0.2182 0.0727 0.1091]
PageRank : [0.2364 0.3636 0.2182 0.0727 0.1091]
fois 55  : [13. 20. 12.  4.  6.]
```

Deux détails du programme méritent une explication. La fonction `np.linalg.eig` peut renvoyer des nombres complexes lorsque la matrice n'est pas symétrique ; la fonction `np.real` conserve la partie réelle, puisque le vecteur propre cherché est réel. Et comme on ne connaît pas à l'avance la position de la valeur propre 1 dans la liste, on cherche celle qui en est la plus proche.

Le vecteur obtenu est

$$\mathbf{q} = \tfrac{1}{55}\,(13, 20, 12, 4, 6).$$

On vérifie que $A\mathbf{q} = \mathbf{q}$ et que la matrice A est régulière (tous les coefficients de A^5 sont strictement positifs), de sorte que le théorème 4.6 s'applique : quelle que soit la page de départ, l'internaute passe à long terme $\frac{20}{55}$ de son temps sur la page 2, $\frac{13}{55}$ sur la page 1, et ainsi de suite. Les composantes de $\mathbf{q}$ forment le **PageRank** des pages, et le classement par ordre décroissant est : page 2, page 1, page 3, page 5, page 4.

L'équation $\mathbf{q} = A\mathbf{q}$, lue ligne par ligne, révèle le principe de l'algorithme. Sa deuxième ligne s'écrit

$$q_2 = 1 \cdot q_1 + \tfrac{1}{3}\,q_3 + \tfrac{1}{2}\,q_5 = \tfrac{13}{55} + \tfrac{1}{3} \cdot \tfrac{12}{55} + \tfrac{1}{2} \cdot \tfrac{6}{55} = \tfrac{20}{55}.$$

L'importance d'une page est la somme des « votes » qu'elle reçoit des pages qui pointent vers elle ; chaque page partage son propre PageRank également entre tous ses liens sortants. Un lien issu d'une page importante qui contient peu de liens rapporte donc beaucoup : la page 1, qui ne reçoit que deux liens, devance la page 3, qui en reçoit autant, parce que l'un de ses deux liens vient de la page 2, la plus importante. La définition semble circulaire, puisque l'importance d'une page dépend de celle des autres ; l'algèbre linéaire lève cette circularité, en faisant de $\mathbf{q}$ la solution d'un système d'équations.

Note historique 4.3 Le gogol et un vecteur propre à 25 milliards

Le projet de recherche de Larry Page et Sergey Brin sur l'analyse des liens du Web (note historique 4.4) a donné naissance au moteur de recherche Google, décrit dans un article présenté en 1998 [BP98]. Le nom Google est inspiré du terme *googol* (*gogol* en français), qui désigne le nombre 10^{100}. Selon le récit qu'en fait le mathématicien états-unien Edward Kasner dans son livre *Mathematics and the Imagination* (1940), ce mot aurait été inventé par son neveu de neuf ans, Milton Sirotta. L'analyse mathématique de l'algorithme, que Kurt Bryan et Tanya Leise ont popularisée sous le titre « The $25,000,000,000 eigenvector » (le vecteur propre de 25 milliards de dollars) [BL06], repose entièrement sur les notions de ce chapitre.

4.5.2 Les pages sans liens sortants

Le Web réel contient des pages qui ne pointent vers aucune autre page : des documents PDF, des images, des pages en construction. Ajoutons au mini-web précédent une page 6, vers laquelle pointe la page 5, mais qui ne contient aucun lien (figure 4.5).

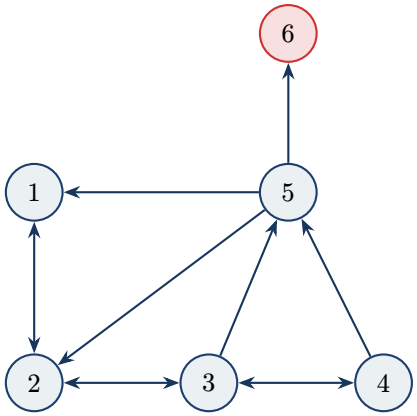


FIGURE 4.5 : Le mini-web 2 : la page 6 n'a aucun lien sortant. Un internaute qui y arrive ne peut plus poursuivre sa promenade en cliquant.

La sixième colonne de la matrice de transition A est alors nulle, et A n'est plus stochastique : la somme de sa sixième colonne vaut 0. Concrètement, la probabilité « s'échappe » du système : un internaute qui arrive sur la page 6 n'a nulle part où aller. Pour corriger ce défaut, on convient que l'internaute qui arrive sur une page sans lien sortant choisit ensuite une page *au hasard* parmi les N pages du Web, chacune avec la probabilité $1/N$. On remplace ainsi chaque colonne nulle de A par une colonne dont tous les coefficients valent $1/N$. Pour le mini-web 2, on obtient

$$B = \begin{pmatrix} 0 & \frac{1}{2} & 0 & 0 & \frac{1}{3} & \frac{1}{6} \\ 1 & 0 & \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{6} \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & \frac{1}{6} \\ 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{6} \\ 0 & 0 & \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{6} \\ 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{6} \end{pmatrix}.$$

Cette correction s'écrit de façon compacte. Soit $\mathbf{E}$ le vecteur de $\mathbb{R}^N$ dont la j^{e} composante vaut 1 si la page j n'a pas de lien sortant, et 0 sinon ; ici, $\mathbf{E} = (0, 0, 0, 0, 0, 1)$. La matrice $\mathbf{e}\,\mathbf{E}^{\intercal}$, produit d'une colonne et d'une ligne, a pour colonnes $\mathbf{e}$ là où $E_j = 1$ et $\mathbf{0}$ ailleurs. Par conséquent,

$$B = A + \frac{1}{N}\,\mathbf{e}\,\mathbf{E}^{\intercal}. \tag{4.2}$$

La matrice B est stochastique et régulière. Son état d'équilibre, que l'on peut calculer avec le programme de la section 4.5.3, est

$$\mathbf{q} = \tfrac{1}{127}\,(27, 42, 27, 10, 15, 6) \approx (0{,}213\,;\ 0{,}331\,;\ 0{,}213\,;\ 0{,}079\,;\ 0{,}118\,;\ 0{,}047).$$

Le classement devient : page 2, pages 1 et 3 à égalité, page 5, page 4 et page 6. L'ajout de la page 6 a peu modifié le classement des cinq premières pages : seule la page 1, qui devançait légèrement la page 3, la rejoint à égalité, en partie parce que la page 5 partage maintenant son vote entre trois liens au lieu de deux.

4.5.3 Les régions piégeantes et la téléportation

Une seconde difficulté est plus sérieuse. Modifions le mini-web 2 en ajoutant une page 7, et des liens de la page 6 vers la page 7 et de la page 7 vers la page 6 (figure 4.6). Il n'y a plus de page sans lien sortant, mais un internaute qui arrive sur la page 6 ne peut plus en sortir : il fait indéfiniment l'aller-retour entre les pages 6 et 7. Les pages 6 et 7 forment une **région piégeante**.

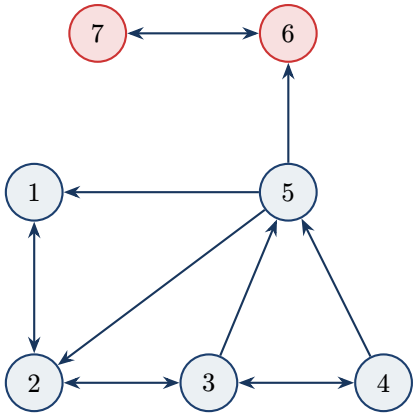


FIGURE 4.6 : Le mini-web 3 : les pages 6 et 7 forment une région piégeante. On peut y entrer, par la page 5, mais on ne peut plus en sortir.

La matrice de transition A de ce mini-web est stochastique, mais elle n'est pas régulière : aucune puissance de A ne permet d'aller de la page 6 vers la page 1. Son unique état d'équilibre est $\mathbf{q} = (0, 0, 0, 0, 0, \frac{1}{2}, \frac{1}{2})$: à long terme, l'internaute est presque sûrement piégé, et les pages 1 à 5 reçoivent toutes un PageRank nul. Un tel classement est inutilisable. De plus, la loi $\mathbf{p}_n$ ne converge pas, puisque l'internaute alterne entre les pages 6 et 7 comme dans l'exemple 4.5 ; on vérifie d'ailleurs que -1 est une valeur propre de A. Sur le Web réel, des régions piégeantes se forment spontanément, mais elles peuvent aussi être créées délibérément, par un ensemble de pages qui se citent mutuellement pour accaparer le PageRank.

La solution retenue par Brin et Page consiste à donner à l'internaute un pouvoir de *téléportation* : à chaque étape, il suit un lien de la page où il se trouve avec une probabilité β, et, avec la probabilité $1-\beta$, il abandonne sa navigation et saute vers une page choisie au hasard parmi les N pages du Web. On peut penser à un internaute qui, de temps en temps, tape une nouvelle adresse dans son navigateur. Le nombre $\beta \in]0, 1[$ s'appelle le **facteur d'amortissement** ; Brin et Page indiquaient utiliser habituellement $\beta = 0{,}85$ [BP98]. La matrice de transition de cette nouvelle promenade est

$$G = \beta B + (1-\beta)\, U, \qquad \text{où} \qquad U = \frac{1}{N}\, \mathbf{e}\, \mathbf{e}^{\mathsf{T}} \tag{4.3}$$

est la matrice $N \times N$ dont tous les coefficients valent $1/N$, et B est la matrice (4.2) qui tient compte des pages sans liens sortants. On reconnaît le « tirage au sort » de l'exemple 4.3, avec la loi uniforme au lieu de $\mathbf{q}$.

Proposition 4.7 **Propriétés de la matrice G**

Si $0 < \beta < 1$, la matrice G est stochastique et tous ses coefficients sont strictement positifs. En particulier, G est régulière, et le théorème 4.6 s'applique : G possède un unique état d'équilibre $\mathbf{q}$, dont toutes les composantes sont strictement positives.

Démonstration. Les matrices B et U sont stochastiques. Chaque coefficient de G est positif, et la somme d'une colonne de G vaut $\beta \cdot 1 + (1-\beta) \cdot 1 = 1$. De plus, chaque coefficient de G est au moins égal à $(1-\beta)/N > 0$. La matrice G est donc régulière, avec $k = 1$ dans la définition 4.6. □

Le vecteur $\mathbf{q}$ est le **PageRank** du Web. Pour le mini-web 3, avec $\beta = 0{,}85$ et $N = 7$, le programme suivant le calcule. La fonction `matrice_google` construit B, puis G, à partir de la liste des liens ; avec `beta=1`, elle ne fait que corriger les pages sans liens sortants, ce qui permet de retrouver le PageRank du mini-web 2 et l'équilibre du mini-web 3 sans téléportation.

Programme 4.7 : La matrice Google et le PageRank des mini-webs 2 et 3. La fonction `matrice_transition` est celle du programme 4.5.

```
def matrice_google(liens, N, beta=0.85):
    A = matrice_transition(liens, N)
    E = (A.sum(axis=0) == 0).astype(float)  # colonnes nulles
    e = np.ones(N)
    B = A + np.outer(e, E) / N              # pages sans sortie
    return beta * B + (1 - beta) * np.outer(e, e) / N

def equilibre(M):
    valeurs, vecteurs = np.linalg.eig(M)
    q = np.real(vecteurs[:, np.argmin(abs(valeurs - 1))])
    return q / q.sum()

liens2 = {1: [2], 2: [1, 3], 3: [2, 4, 5], 4: [3, 5],
          5: [1, 2, 6], 6: []}
G2 = matrice_google(liens2, 6, beta=1)
print("mini-web 2 :", equilibre(G2))

liens3 = {1: [2], 2: [1, 3], 3: [2, 4, 5], 4: [3, 5],
          5: [1, 2, 6], 6: [7], 7: [6]}
G3 = matrice_google(liens3, 7, beta=1)
print("sans téléportation :", equilibre(G3))
print("beta = 0,85 :", equilibre(matrice_google(liens3, 7)))
```

Programme 4.8 : Sortie.

```
mini-web 2 : [0.2126 0.3307 0.2126 0.0787 0.1181 0.0472]
sans téléportation : [-0.  -0.  -0.  -0.  -0.   0.5  0.5]
beta = 0,85 : [0.1216 0.1823 0.1228 0.0562 0.0801 0.2247 0.2124]
```

La fonction `np.outer(u, v)` calcule le produit $\mathbf{u}\,\mathbf{v}^{\mathsf{T}}$ d'une colonne par une ligne. Les valeurs `-0.` de la deuxième ligne sont des nombres de l'ordre de 10^{-16}, négatifs à cause des erreurs d'arrondi, que l'affichage arrondit à zéro.

Avec la téléportation, toutes les pages reçoivent un PageRank positif, et le classement est : page 6, page 7, page 2, page 3, page 1, page 5, page 4. Les pages de la région piégeante restent en tête, puisque l'internaute y reste longtemps chaque fois qu'il y entre, mais elles ne monopolisent plus le classement. Le choix de β règle ce compromis. Le tableau 4.3 montre que, plus β est proche de 1, plus le modèle suit fidèlement la structure des liens, mais plus la région piégeante accapare le PageRank ; plus β est petit, plus le classement se rapproche de la loi uniforme et perd son information.

TABLE 4.3 : Le PageRank du mini-web 3 pour trois valeurs du facteur d'amortissement β. La dernière colonne donne la part totale de la région piégeante formée des pages 6 et 7.

β	q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_6+q_7
0,5	0,137	0,183	0,141	0,095	0,119	0,169	0,156	0,325
0,85	0,122	0,182	0,123	0,056	0,080	0,225	0,212	0,437
0,99	0,031	0,048	0,031	0,012	0,017	0,432	0,429	0,861

4.5.4 La matrice Google

Rassemblons les deux corrections. Pour un Web de N pages, de matrice de transition A, la **matrice Google** est

$$G = \beta B + (1-\beta)\,U = \beta A + \frac{\beta}{N}\,\mathbf{e}\,\mathbf{E}^{\mathsf{T}} + \frac{1-\beta}{N}\,\mathbf{e}\,\mathbf{e}^{\mathsf{T}} = \beta A + \frac{1}{N}\,\mathbf{e}\,\big(\beta\mathbf{E}^{\mathsf{T}} + (1-\beta)\,\mathbf{e}^{\mathsf{T}}\big),$$

et le PageRank est l'unique vecteur de probabilités $\mathbf{q}$ tel que $G\mathbf{q} = \mathbf{q}$. Pour le Web réel, N se compte en milliards : il est hors de question de stocker la matrice G, qui contiendrait plus de 10^{18} coefficients, tous non nuls. La formule précédente montre qu'il n'est pas nécessaire de le faire : il suffit de connaître la matrice A, le vecteur $\mathbf{E}$ et le nombre β. Or A est une matrice **creuse** : une page contient en général quelques dizaines de liens au plus, de sorte que chaque colonne de A ne compte qu'un petit nombre de coefficients non nuls parmi des milliards. On stocke alors seulement les coefficients non nuls et leurs positions, ce qui revient à stocker la liste des liens. Il reste à calculer $\mathbf{q}$ sans jamais former G : c'est l'objet de la section suivante.

4.6 Calculer le PageRank : la méthode des puissances

Pour une matrice de quelques milliards de lignes, la fonction `np.linalg.eig` est inutilisable : elle calcule *toutes* les valeurs propres et tous les vecteurs propres, alors que nous n'en voulons qu'un, et elle travaille sur une matrice pleine. Le théorème 4.6 suggère une méthode beaucoup plus simple : partir d'un vecteur de probabilités quelconque et appliquer G de façon répétée, puisque $G^n\mathbf{x}_0 \to \mathbf{q}$. C'est la **méthode des puissances**.

Algorithme 4.1 **Méthode des puissances pour le PageRank**

Données : la matrice creuse A, le vecteur $\mathbf{E}$, le facteur β et une tolérance $\varepsilon > 0$.

1. Poser $\mathbf{w}_0 = \frac{1}{N}\,\mathbf{e}$ (la loi uniforme).
2. Pour $k = 1, 2, \ldots$, calculer
$$\mathbf{w}_k = G\mathbf{w}_{k-1} = \beta A\mathbf{w}_{k-1} + \frac{\beta\, s_{k-1} + 1 - \beta}{N}\,\mathbf{e}, \qquad \text{où } s_{k-1} = \mathbf{E}^{\mathsf{T}}\mathbf{w}_{k-1}.$$
3. S'arrêter lorsque $d_1(\mathbf{w}_k, \mathbf{w}_{k-1}) < \varepsilon$, et renvoyer $\mathbf{w}_k$.

La formule de l'étape 2 provient de l'expression de G à la section 4.5.4 : pour un vecteur de probabilités $\mathbf{w}$, on a $\mathbf{e}^{\mathsf{T}}\mathbf{w} = 1$, donc

$$G\mathbf{w} = \beta A\mathbf{w} + \frac{1}{N}\,\mathbf{e}\,\big(\beta\,\mathbf{E}^{\mathsf{T}}\mathbf{w} + (1-\beta)\,\mathbf{e}^{\mathsf{T}}\mathbf{w}\big) = \beta A\mathbf{w} + \frac{\beta\,\mathbf{E}^{\mathsf{T}}\mathbf{w} + 1 - \beta}{N}\,\mathbf{e}.$$

Le nombre $s = \mathbf{E}^{\mathsf{T}}\mathbf{w}$ est simplement la somme des composantes de $\mathbf{w}$ qui correspondent aux pages sans liens sortants. Le coût d'une itération est donc essentiellement celui du

produit $A\mathbf{w}$, proportionnel au nombre de liens du Web plutôt qu'au carré du nombre de pages. La distance d_1 utilisée dans le critère d'arrêt est définie par

$$d_1(\mathbf{x}, \mathbf{y}) = |x_1 - y_1| + |x_2 - y_2| + \cdots + |x_N - y_N|;$$

c'est la taxi-distance de l'exemple 2.2, étendue à $\mathbb{R}^N$. Le programme suivant met en œuvre l'algorithme.

Programme 4.9 : La méthode des puissances, appliquée au mini-web 3.

```
def pagerank(liens, N, beta=0.85, tol=1e-8):
    """Méthode des puissances, sans construire G."""
    A = matrice_transition(liens, N)    # creuse, en pratique
    sans_sortie = (A.sum(axis=0) == 0)  # le vecteur E
    w = np.ones(N) / N                  # départ : loi uniforme
    for k in range(1, 1000):
        s = w[sans_sortie].sum()        # le produit E^T w
        w_suiv = beta * (A @ w) + (beta * s + 1 - beta) / N
        if np.abs(w_suiv - w).sum() < tol:   # d_1(w_suiv, w)
            return w_suiv, k
        w = w_suiv

liens3 = {1: [2], 2: [1, 3], 3: [2, 4, 5], 4: [3, 5],
          5: [1, 2, 6], 6: [7], 7: [6]}
q, k = pagerank(liens3, 7)
print(k, "itérations")
print(q)
```

Programme 4.10 : Sortie.

```
93 itérations
[0.1216 0.1823 0.1228 0.0562 0.0801 0.2247 0.2124]
```

On retrouve le PageRank obtenu à la section 4.5.3. Dans le programme, l'addition d'un vecteur et d'un nombre ajoute ce nombre à chaque composante, ce qui réalise le terme $\frac{\beta s + 1 - \beta}{N}\,\mathbf{e}$ sans construire le vecteur $\mathbf{e}$.

4.6.1 Pourquoi la méthode converge : retour au théorème de Banach

Le théorème 4.6 garantit la convergence, mais il ne dit pas combien d'itérations sont nécessaires, et nous ne l'avons démontré que pour les matrices diagonalisables. Le théorème du point fixe de Banach du chapitre 2 fournit une démonstration complète, assortie d'une estimation de l'erreur. Comme l'annonçait la note historique 2.4, le PageRank est le point fixe d'une contraction.

Théorème 4.8 **La matrice Google est une contraction**

Soit Δ l'ensemble des vecteurs de probabilités de $\mathbb{R}^N$, muni de la distance d_1. Pour tous $\mathbf{x}, \mathbf{y} \in \Delta$,

$$d_1(G\mathbf{x}, G\mathbf{y}) \leq \beta\, d_1(\mathbf{x}, \mathbf{y}).$$

L'application $\mathbf{x} \mapsto G\mathbf{x}$ est donc une contraction de rapport β sur (Δ, d_1).

Démonstration. Posons $\mathbf{z} = \mathbf{x} - \mathbf{y}$. Les sommes des composantes de $\mathbf{x}$ et de $\mathbf{y}$ valent

toutes deux 1, donc $\mathbf{e}^{\mathsf{T}}\mathbf{z} = 0$, et

$$G\mathbf{z} = \beta B\mathbf{z} + \frac{1-\beta}{N}\,\mathbf{e}\,(\mathbf{e}^{\mathsf{T}}\mathbf{z}) = \beta B\mathbf{z}.$$

Il reste à voir que $d_1(B\mathbf{z}, \mathbf{0}) \leq d_1(\mathbf{z}, \mathbf{0})$. Par l'inégalité triangulaire dans $\mathbb{R}$, puis en échangeant l'ordre des sommes,

$$\sum_{i=1}^{N} \Big| \sum_{j=1}^{N} b_{ij} z_j \Big| \leq \sum_{i=1}^{N} \sum_{j=1}^{N} b_{ij}\,|z_j| = \sum_{j=1}^{N} |z_j| \sum_{i=1}^{N} b_{ij} = \sum_{j=1}^{N} |z_j|,$$

puisque les coefficients de B sont positifs et que la somme de chaque colonne de B vaut 1. Ainsi $d_1(G\mathbf{x}, G\mathbf{y}) = \beta\, d_1(B\mathbf{z}, \mathbf{0}) \leq \beta\, d_1(\mathbf{x}, \mathbf{y})$. □

L'espace métrique (Δ, d_1) est complet : si une suite de vecteurs de Δ est de Cauchy pour d_1, chacune de ses composantes est une suite de Cauchy de nombres réels, donc converge, et la limite a encore des composantes positives de somme 1. Par le théorème de Banach, sous sa forme générale (remarque 2.11), l'application $\mathbf{x} \mapsto G\mathbf{x}$ possède un unique point fixe $\mathbf{q}$ dans Δ, et la méthode des puissances converge vers $\mathbf{q}$ quel que soit le vecteur de départ. On retrouve ainsi l'existence et l'unicité du PageRank sans faire appel au théorème de Perron–Frobenius. La majoration de l'erreur (proposition 2.7) devient

$$d_1(\mathbf{w}_k, \mathbf{q}) \leq \frac{\beta^k}{1-\beta}\, d_1(\mathbf{w}_0, \mathbf{w}_1).$$

En l'appliquant à partir de $\mathbf{w}_{k-1}$ plutôt que de $\mathbf{w}_0$, avec un seul pas, on obtient aussi $d_1(\mathbf{w}_k, \mathbf{q}) \leq \frac{\beta}{1-\beta}\, d_1(\mathbf{w}_{k-1}, \mathbf{w}_k)$: lorsque le critère d'arrêt de l'algorithme 4.1 est satisfait, l'erreur est inférieure à $\frac{0{,}85}{0{,}15}\,\varepsilon \approx 5{,}7\,\varepsilon$.

Le nombre d'itérations nécessaires ne dépend donc pas de la taille du Web, mais seulement de β et de la précision voulue. Comme deux vecteurs de probabilités sont à une distance d_1 d'au plus 2, on a $d_1(\mathbf{w}_k, \mathbf{q}) \leq \beta^k\, d_1(\mathbf{w}_0, \mathbf{q}) \leq 2\beta^k$; pour garantir une erreur inférieure à 10^{-8} avec $\beta = 0{,}85$, il suffit que $2\,(0{,}85)^k < 10^{-8}$, c'est-à-dire $k \geq 118$. Cette majoration est pessimiste : le programme précédent s'est arrêté après 93 itérations. Page et ses collaborateurs rapportaient qu'une cinquantaine d'itérations suffisaient en pratique pour obtenir un classement stable sur une base de plusieurs centaines de millions de liens [Pag+99].

Remarque 4.3 **Le rôle de la seconde valeur propre**

La vitesse de convergence de la méthode des puissances est gouvernée, comme à la section 4.4, par le plus grand module $|\lambda_2|$ des valeurs propres de G autres que 1. On démontre que $|\lambda_2| \leq \beta$ pour toute matrice Google [LM06], ce qui est cohérent avec le théorème 4.8. Pour le mini-web 3, l'égalité est atteinte : G possède une valeur propre égale à $-0{,}85$, héritée de la valeur propre -1 de la matrice A, qui traduit l'aller-retour entre les pages 6 et 7. Le facteur d'amortissement joue donc un double rôle : il règle le compromis entre fidélité à la structure des liens et poids de la téléportation (tableau 4.3), et il fixe la vitesse de convergence. Choisir β proche de 1 rend le calcul plus lent : avec $\beta = 0{,}99$, il faudrait environ 1900 itérations pour garantir la même précision.

Pour conclure, rappelons que PageRank n'est qu'un des nombreux critères qu'utilisent les moteurs de recherche actuels, et que ses détails de mise en œuvre chez Google n'ont jamais été entièrement publiés. L'idée qu'il incarne a toutefois connu de nombreuses

applications en dehors du Web : classer des articles scientifiques selon les citations qu'ils reçoivent, des équipes sportives selon leurs victoires, ou des espèces selon leur rôle dans un réseau alimentaire. Chaque fois, il s'agit de mesurer l'importance d'un élément d'un réseau par l'importance des éléments qui pointent vers lui, et le calcul se ramène à la recherche d'un vecteur propre associé à la valeur propre 1.

Pour la classe 4.4 Classer, une activité pour la classe

Le principe de PageRank peut être exploré sans matrices. On présente aux élèves un petit réseau de quatre ou cinq pages, dessiné au tableau, et on leur demande d'abord de prédire un classement : la plupart compteront les liens entrants. On simule ensuite l'internaute aléatoire avec un dé, chaque élève ou chaque équipe notant les pages visitées pendant une trentaine de clics, et l'on compare les fréquences obtenues à la prédiction. Un réseau bien choisi, comme celui de la figure 4.1, où une page qui reçoit peu de liens devance une page qui en reçoit autant, provoque une discussion sur ce que signifie « être important » dans un réseau. Dans la séquence Technico-sciences, on peut prolonger l'activité par le calcul de $A\mathbf{p}$ à l'aide d'un tableur, en lien avec la multiplication matricielle [Min16].

Note historique 4.4 Larry Page



Larry Page (né en 1973), en 2014. Photo : Steve Jurvetson, CC BY 2.0, *Wikimedia Commons*.

Lawrence (Larry) Page est né en 1973 à East Lansing, au Michigan. Ses deux parents enseignaient l'informatique à l'Université d'État du Michigan, et il grandit entouré d'ordinateurs. Après des études de génie informatique à l'Université du Michigan, il entra en 1995 au doctorat en informatique de l'Université Stanford, sous la direction de Terry Winograd. C'est là qu'il rencontra Sergey Brin, un autre doctorant, né à Moscou la même année que lui.

Page s'intéressait à la structure des liens du Web, plutôt qu'à son contenu. Il remarqua l'analogie avec les citations entre articles scientifiques : un article souvent cité, surtout par des articles eux-mêmes importants, est un article important. Pour appliquer cette idée au Web, il fallait d'abord en télécharger les liens, ce que fit le robot du projet *BackRub* à partir de 1996 ; le titre de l'article de 1999 sur PageRank, « The PageRank Citation Ranking » [Pag+99], garde la trace de cette analogie. Le nom de l'algorithme fait d'ailleurs un jeu de mots entre le nom de famille de son auteur et la *page* web. Le brevet sur la méthode, déposé par l'Université Stanford, porte le nom de Page comme inventeur ; Stanford en accorda une licence exclusive à Google.

En septembre 1998, Page et Brin fondèrent l'entreprise Google, dans un garage de Menlo Park, en Californie ; Page abandonna son doctorat pour s'y consacrer. Il en fut le premier président-directeur général, jusqu'en 2001, puis de nouveau de 2011 à 2015, avant de diriger Alphabet, la société mère de Google, jusqu'en 2019. Son parcours illustre la distance parfois très courte entre une idée mathématique et ses applications : à l'origine de l'une des plus grandes entreprises du monde se trouve un projet de doctorat, dont le cœur est le calcul d'un vecteur propre associé à la

valeur propre 1.

Exercices

Exercice 4.1. Florence a trouvé que $\mathbf{v} = (1, -2, 6)$ est un vecteur propre associé à une certaine valeur propre λ d'une matrice A. En utilisant un logiciel pour vérifier son résultat, elle a plutôt obtenu le vecteur $\mathbf{w} = (-5, 10, -30)$. Ne trouvant pas son erreur, elle a demandé de l'aide à son enseignante, qui lui a expliqué que le vecteur obtenu par le logiciel est aussi une bonne réponse. Expliquer pourquoi.

Exercice 4.2. Montrer que si $\mathbf{v}$ est un vecteur propre d'une matrice A associé à une valeur propre λ, alors $\mu\mathbf{v}$ est lui aussi un vecteur propre de A associé à λ, quel que soit le nombre réel $\mu \neq 0$. Pourquoi faut-il exclure $\mu = 0$?

Exercice 4.3. Pour chacune des matrices suivantes, déterminer le polynôme caractéristique $p(\lambda) = \det(A - \lambda I)$, les valeurs propres, et les vecteurs propres associés à chaque valeur propre.

a) $\begin{pmatrix} 6 & 14 & 9 \\ 0 & 3 & 0 \\ 0 & -10 & -3 \end{pmatrix}$

b) $\begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \\ 4 & 4 & 4 \end{pmatrix}$

c) $\begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}$, où $a, b \in \mathbb{R}$.

d) $\begin{pmatrix} 1 & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & \frac{1}{3} \\ 0 & \frac{1}{2} & \frac{1}{3} \end{pmatrix}$. Cette matrice est-elle stochastique ? Interpréter l'état 1 dans la chaîne de Markov correspondante.

Exercice 4.4. Soit P une matrice carrée ayant λ comme valeur propre et $\mathbf{v}$ comme vecteur propre associé à λ. Montrer par récurrence que, pour tout entier $n \geq 1$, λ^n est une valeur propre de P^n et $\mathbf{v}$ est un vecteur propre de P^n associé à λ^n.

Exercice 4.5. Montrer que deux vecteurs de probabilités *distincts* ne peuvent pas être colinéaires. En déduire que, si 1 est une valeur propre simple d'une matrice stochastique, celle-ci possède au plus un état d'équilibre.

Exercice 4.6. Une matrice M possède un vecteur propre $\mathbf{v}$ qui est un vecteur de probabilités. La matrice M est-elle nécessairement stochastique ? Justifier.

Exercice 4.7. Soient M une matrice stochastique et $\mathbf{v}$ un vecteur de probabilités. Montrer que $M\mathbf{v}$ est aussi un vecteur de probabilités (proposition 4.3).

Exercice 4.8. Soit M une matrice stochastique de taille N, et notons $\|\mathbf{v}\|_1 = |v_1| + \cdots + |v_N|$.

a) Montrer que $\|M\mathbf{v}\|_1 \leq \|\mathbf{v}\|_1$ pour tout vecteur $\mathbf{v} \in \mathbb{R}^N$.

b) En déduire que toute valeur propre réelle λ de M vérifie $|\lambda| \leq 1$.

c) Montrer que si un vecteur propre de M est un vecteur de probabilités, la valeur propre associée vaut 1.

Exercice 4.9. Écrire une fonction Python `vers_probabilites(v)` qui prend un vecteur **v** et renvoie le vecteur de probabilités colinéaire à **v**, lorsqu'il existe. À quelle condition sur **v** existe-t-il ? Est-il unique ? Tester la fonction sur les vecteurs propres renvoyés par `np.linalg.eig` pour la matrice M du modèle météorologique.

Exercice 4.10. On reprend le modèle météorologique à mémoire de deux jours de la section 4.3.3, de matrice M_2.

a) Vérifier que $\mathbf{q} = \frac{1}{32}(21, 3, 3, 5)$ est l'état d'équilibre de M_2.

b) Il a plu hier et il fait beau aujourd'hui. Quelle est la probabilité qu'il fasse beau après-demain ?

c) À long terme, quelle est la probabilité de deux jours de pluie consécutifs ? Comparer avec le modèle à deux états de la section 4.2.

Exercice 4.11. Dans une ville, on distingue trois types de véhicules : les automobiles de type berline ou coupé (A), les véhicules utilitaires sport (V) et les camionnettes (C). Chaque année, on observe les transitions suivantes :

— 3 % des propriétaires d'un A l'échangent pour un V, et 1 % l'échangent pour un C ;

— 2 % des propriétaires d'un V l'échangent pour un A, et 3 % l'échangent pour un C ;

— 5 % des propriétaires d'un C l'échangent pour un V, et 2 % l'échangent pour un A.

Actuellement, les proportions sont de 65 % pour A, 30 % pour V et 5 % pour C. En supposant que le nombre de véhicules reste constant, déterminer les proportions de chaque type dans 1 an, dans 5 ans, dans 10 ans et à long terme.

Exercice 4.12. Doudou, le hamster paresseux, connaît trois endroits dans sa cage : les copeaux où il dort, la mangeoire où il mange, et la roue où il fait de l'exercice. Ses journées se ressemblent, et son activité peut être modélisée par une chaîne de Markov, en observant l'endroit où il se trouve à intervalles réguliers (par exemple chaque minute) [1].

— Quand il dort, il a $\frac{9}{10}$ des chances de ne pas se réveiller à l'observation suivante. Lorsqu'il se réveille, il a une chance sur deux d'aller manger et une chance sur deux d'aller faire de l'exercice.

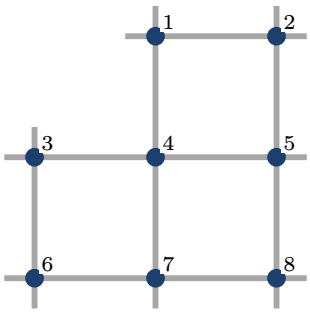
— Après avoir mangé, il a $\frac{7}{10}$ des chances de retourner dormir et $\frac{3}{10}$ des chances d'aller courir.

— Après avoir couru, il a $\frac{8}{10}$ des chances de retourner dormir ; sinon, il continue à courir.

Combien de temps Doudou passe-t-il à dormir, en moyenne, dans une journée ?

Exercice 4.13. Un policier contrôle la circulation aux huit intersections illustrées ci-dessous. Lorsqu'il arrive à une intersection, il y demeure une heure, puis il décide au hasard d'aller à une intersection voisine ou de rester où il est, chaque possibilité ayant la même probabilité d'être choisie. Par exemple, s'il est à l'intersection 5, il choisit d'aller à 2, 4, 5 ou 8, avec une probabilité de $\frac{1}{4}$ pour chacune. À la fin de sa journée de travail, il choisit de la même façon le point de départ du lendemain.

1. Exemple tiré de l'article « Chaîne de Markov » de Wikipédia, qui utilise la convention des lignes (remarque 4.1).

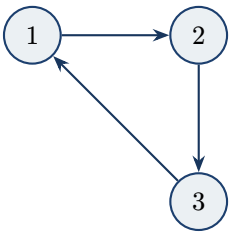


Déterminer le pourcentage de son temps de travail que le policier passe à chacune des intersections.

Exercice 4.14. La matrice suivante est la matrice de transition d'une chaîne de Markov associée à un mini-web. Dessiner ce mini-web.

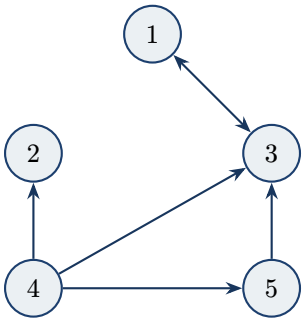
$$\begin{pmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{3} & 0 & \frac{1}{2} & 0 \end{pmatrix}$$

Exercice 4.15. Considérons le mini-web suivant.



a) Sans faire de calculs, déterminer le PageRank de ce mini-web (sans téléportation).
b) Peut-on ajouter d'autres liens tout en conservant le même PageRank ?
c) La loi $\mathbf{p}_n$ d'un internaute qui part de la page 1 converge-t-elle ? Comment interpréter alors le PageRank ?

Exercice 4.16. Soit $\mathcal{W}$ le mini-web suivant.

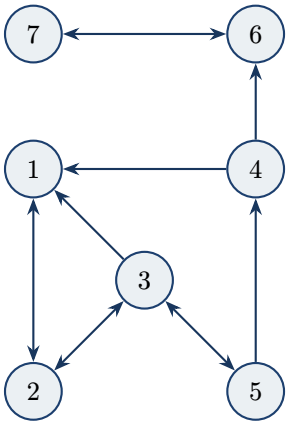


a) Si un internaute commence sa promenade à la page 5, quelle proportion de son temps passera-t-il sur chaque page ?
b) Que se passe-t-il s'il commence à la page 2 ?
c) Déterminer la matrice de transition M de $\mathcal{W}$.
d) Trouver un vecteur $\mathbf{v}$ non nul tel que $M\mathbf{v} = \mathbf{v}$.
e) Trouver le polynôme caractéristique de M, puis ses valeurs propres.

f) Modifier la matrice pour tenir compte des pages sans liens sortants. On appellera N cette matrice.

g) Calculer la matrice Google G avec une probabilité de téléportation de $\frac{1}{10}$, puis utiliser G pour déterminer le PageRank.

Exercice 4.17. Considérons le mini-web $\mathcal{W}$ ci-dessous.



a) Trouver la matrice de transition de $\mathcal{W}$. Contient-il des pages sans liens sortants ? des régions piégeantes ?

b) Trouver la matrice Google de $\mathcal{W}$, avec une probabilité de téléportation de $0{,}15$.

c) À l'aide d'un outil informatique, trouver le PageRank de $\mathcal{W}$, et classer les pages.

Exercice 4.18. Considérons un mini-web $\mathcal{W}$ de $N \geq 2$ pages, de matrice de transition A, et les vecteurs $\mathbf{e} = (1, 1, \dots, 1)$ et $\mathbf{E} = (E_1, \dots, E_N)$, où $E_j = 1$ si la page j n'a pas de liens sortants et $E_j = 0$ sinon.

a) À quoi correspond la matrice $\frac{1}{N}\,\mathbf{e}\,\mathbf{E}^\intercal$?

b) Exprimer la matrice B qui tient compte des pages sans liens sortants.

c) Montrer que la matrice Google, avec un facteur d'amortissement β, peut s'écrire

$$G = \beta A + \frac{1}{N}\,\mathbf{e}\,(\beta \mathbf{E}^\intercal + (1-\beta)\,\mathbf{e}^\intercal).$$

d) Pourquoi A est-elle une matrice creuse lorsque $\mathcal{W}$ est le Web réel ? Pourquoi, si $0 < \beta < 1$, tous les coefficients de G sont-ils dans l'intervalle $]0, 1[$?

Exercice 4.19. Les chaînes de Markov ne figurent pas au programme du secondaire, mais plusieurs des notions qu'elles mobilisent y sont étudiées.

a) Relever, dans la progression des apprentissages [Min16], trois notions du secondaire qui interviennent dans ce chapitre, et préciser comment chacune y est utilisée.

b) Concevoir les grandes lignes d'une activité d'environ une période, destinée à des élèves du deuxième cycle, fondée sur le modèle météorologique ou sur l'internaute aléatoire. Quelle conception erronée du hasard l'activité permet-elle d'aborder ?

c) Le programme de mathématique vise aussi à développer le jugement critique des élèves. En quoi l'étude de PageRank peut-elle y contribuer ?

5 L'analyse en composantes principales

Un tableau de données n'est souvent rien d'autre qu'une liste de points. Si l'on note, pour chaque élève d'une classe, son résultat en mathématiques et son résultat en sciences, chaque élève devient un point du plan, et la classe entière, un *nuage de points*. Avec trois mesures par élève, le nuage vit dans l'espace $\mathbb{R}^3$; avec les dix épreuves d'un décathlon, dans $\mathbb{R}^{10}$; avec les centaines de milliers de marqueurs génétiques d'une personne, dans un espace de dimension gigantesque. Dans le plan, un coup d'œil suffit pour repérer une tendance, des groupes ou un point isolé. Mais nos yeux s'arrêtent à la dimension 3, et encore : sur une page ou un écran, nous ne voyons jamais qu'une *projection* de l'espace sur un plan.

Pensons à la photographie d'un objet. Une théière photographiée de dessus n'est qu'un disque ; photographiée de côté, on reconnaît immédiatement son bec et son anse. Les deux photos sont des projections du même objet, mais l'une est beaucoup plus informative que l'autre. L'*analyse en composantes principales* (ACP) répond à la question correspondante pour un nuage de points : *sous quel angle faut-il le regarder pour en voir le plus possible ?* Elle cherche la droite, puis le plan, puis plus généralement le sous-espace de petite dimension sur lequel la projection du nuage déforme le moins les données.

Nous verrons que la réponse est donnée par les vecteurs propres d'une matrice symétrique, la *matrice de covariance* des données. Ce chapitre réunit donc deux sujets qui semblent éloignés, la statistique descriptive du secondaire et les vecteurs propres de l'annexe D. Nous commencerons très modestement, par un nuage de six points du plan, et nous chercherons la droite qui le représente le mieux ; la projection orthogonale (définition D.4) et le théorème de Pythagore feront l'essentiel du travail. Nous passerons ensuite à un nuage de l'espace, formé de mesures prises sur des manchots de l'Antarctique, et nous chercherons le meilleur plan sur lequel le projeter. Nous pourrons alors définir l'ACP en toute généralité, en faire un bref historique et l'appliquer, avec Python, à des données où aucun dessin direct n'est possible.

5.1 Un peu de statistique descriptive

5.1.1 Individus, variables et tableau de données

En statistique, on appelle **individus** les objets sur lesquels on prend des mesures (des élèves, des manchots, des athlètes, des villes), et **variables** les grandeurs mesurées (une note, une longueur, un temps de course, une température). Lorsqu'on mesure p variables sur n individus, on range les résultats dans un **tableau de données** à n lignes et p colonnes : la ligne i décrit l'individu i, et la colonne j contient les valeurs de la variable j. Chaque individu est ainsi représenté par un point de $\mathbb{R}^p$.

Le tableau 5.1 nous servira d'exemple tout au long des deux premières sections. Il donne les résultats (sur 100) de six élèves, désignés par les lettres A à F, à un examen de

mathématiques et à un examen de sciences. Ici, $n = 6$ et $p = 2$: le nuage est formé de six points du plan (figure 5.1).

TABLE 5.1 : Les résultats de six élèves (données fictives), un élève par ligne. Les deux premières colonnes de résultats forment le tableau de données ; les deux dernières donnent les écarts à la moyenne, qui serviront à la section 5.2.

Élève	Mathématiques x	Sciences y	Écart $x - \bar{x}$	Écart $y - \bar{y}$
A	60	72	-10	-3
B	63	66	-7	-9
C	67	67	-3	-8
D	75	84	5	9
E	76	82	6	7
F	79	79	9	4
Moyenne	70	75	0	0

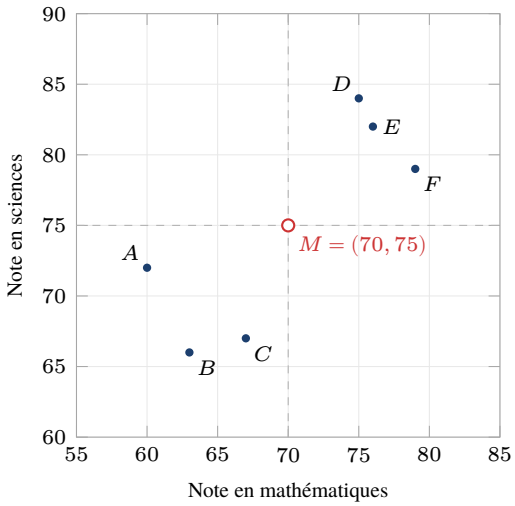


FIGURE 5.1 : Le nuage de points du tableau 5.1. Les élèves forts en mathématiques ont tendance à l'être aussi en sciences : le nuage s'allonge selon une direction oblique. Le point moyen M est le centre de gravité du nuage.

5.1.2 Moyenne, variance et écart type

Définition 5.1 **Moyenne, variance, écart type**

Soit $x_1, \dots, x_n$ les valeurs d'une variable x sur n individus. La **moyenne** de x est

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

La **variance** de x est la moyenne des carrés des écarts à la moyenne, et l'**écart type** est sa racine carrée :

$$s_x^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2, \qquad s_x = \sqrt{s_x^2}.$$

Le point $M = (\bar{x}_1, \dots, \bar{x}_p)$ dont les coordonnées sont les moyennes des p variables s'appelle le **point moyen** (ou centre de gravité) du nuage.

La variance mesure la *dispersion* des valeurs autour de leur moyenne : elle est nulle si toutes les valeurs sont égales, et elle est d'autant plus grande que les valeurs sont étalées. On utilise souvent l'écart type plutôt que la variance, parce qu'il s'exprime dans la même unité que les données. Pour les notes de mathématiques du tableau 5.1,

$$s_x^2 = \frac{(-10)^2 + (-7)^2 + (-3)^2 + 5^2 + 6^2 + 9^2}{6} = \frac{300}{6} = 50, \qquad s_x = \sqrt{50} \approx 7{,}07.$$

Le même calcul donne $s_y^2 = \frac{9+81+64+81+49+16}{6} = 50$ pour les sciences : les deux matières ont, par construction, la même dispersion.

Il est souvent commode de remplacer chaque valeur par son écart à la moyenne, $x_i - \bar{x}$. On dit alors que l'on **centre** les données ; géométriquement, on déplace l'origine au point moyen M. Les données centrées ont une moyenne nulle, et leur variance est la même que celle des données de départ, puisqu'une translation ne change pas la dispersion.

Remarque 5.1 **Variable statistique ou variable aléatoire ?**

Le mot « variable » n'a pas ici le sens qu'il avait au chapitre 4. Une variable aléatoire X est décrite par une loi de probabilité, et sa variance, que l'on note $\mathrm{Var}(X)$ avec des majuscules, se calcule à partir de cette loi : c'est une quantité théorique. Dans ce chapitre, x est une *variable statistique* : une simple colonne de nombres $x_1, \dots, x_n$ que l'on a observés. Ces données peuvent très bien provenir d'un processus aléatoire, par exemple d'un tirage au hasard de personnes dans une population. Une fois recueillies, cependant, leur moyenne et leur variance s_x^2 se calculent directement à partir du tableau : ces calculs descriptifs ne demandent aucune hypothèse sur une loi de probabilité sous-jacente. Pour ne pas confondre les deux notions, nous réservons les minuscules et les notations $\bar{x}$, s_x^2 et s_{xy} aux données observées, et nous n'écrirons pas $\mathrm{Var}(x)$.

Les deux notions sont tout de même liées, à condition de bien choisir la variable aléatoire. Choisissons au hasard l'un des n individus du tableau, chacun avec la probabilité $\frac{1}{n}$, et notons X la valeur de la variable pour cet individu. La loi de X est construite à partir du tableau : X prend les valeurs $x_1, \dots, x_n$, chacune avec la probabilité $\frac{1}{n}$. Son espérance μ et sa variance σ^2 restent des quantités théoriques, calculées à partir de cette loi, mais le calcul redonne exactement les statistiques du tableau :

$$\mu = \mathbb{E}(X) = \sum_{i=1}^{n} \frac{1}{n}\, x_i = \bar{x}, \qquad \sigma^2 = \mathrm{Var}(X) = \sum_{i=1}^{n} \frac{1}{n}\, (x_i - \mu)^2 = s_x^2.$$

Les n individus jouent ainsi le rôle d'une population entière, dont $\bar{x}$ et s_x^2 sont les paramètres. C'est pourquoi les calculatrices notent σ_x l'écart type obtenu en divisant par n.

Cette égalité ne vaut que pour cette variable X particulière. Si les n individus forment un échantillon tiré d'une population plus vaste, la variable aléatoire qui décrit la mesure d'un individu pris au hasard dans cette population a une variance inconnue, qui ne coïncide pas avec s_x^2 : les données permettent seulement de l'estimer (remarque 5.2).

5.1.3 Covariance et corrélation

La variance décrit une variable à la fois. Pour décrire la façon dont *deux* variables varient ensemble, on multiplie leurs écarts à la moyenne.

Définition 5.2 **Covariance, coefficient de corrélation**

La **covariance** de deux variables x et y, mesurées sur les mêmes n individus, est

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Si s_x et s_y sont non nuls, le **coefficient de corrélation linéaire** de x et de y est

$$r_{xy} = \frac{s_{xy}}{s_x\, s_y}.$$

Le signe de chaque produit $(x_i - \bar{x})(y_i - \bar{y})$ dépend du quadrant, délimité par les droites pointillées de la figure 5.1, dans lequel se trouve le point i : il est positif en haut à droite et en bas à gauche, négatif dans les deux autres quadrants. Une covariance positive signale donc que les deux variables ont tendance à être grandes en même temps, et petites en même temps. Dans notre exemple, tous les points sont dans les quadrants « positifs », et

$$s_{xy} = \frac{(-10)(-3) + (-7)(-9) + (-3)(-8) + 5\cdot 9 + 6\cdot 7 + 9\cdot 4}{6} = \frac{240}{6} = 40$$

et

$$r_{xy} = \frac{40}{\sqrt{50}\,\sqrt{50}} = 0{,}8.$$

Le coefficient de corrélation a une interprétation géométrique simple. Si l'on range les écarts à la moyenne des deux variables dans deux vecteurs de $\mathbb{R}^n$,

$$\mathbf{a} = (x_1 - \bar{x}, \dots, x_n - \bar{x}) \quad \text{et} \quad \mathbf{b} = (y_1 - \bar{y}, \dots, y_n - \bar{y}),$$

alors $s_{xy} = \frac{1}{n}\,\mathbf{a}\cdot\mathbf{b}$, $s_x = \frac{1}{\sqrt{n}}\|\mathbf{a}\|$ et $s_y = \frac{1}{\sqrt{n}}\|\mathbf{b}\|$, de sorte que, d'après la formule (D.1),

$$r_{xy} = \frac{\mathbf{a}\cdot\mathbf{b}}{\|\mathbf{a}\|\,\|\mathbf{b}\|} = \cos\theta,$$

où θ est l'angle entre $\mathbf{a}$ et $\mathbf{b}$. Comme tout cosinus, r_{xy} vérifie donc $-1 \leq r_{xy} \leq 1$. On a $r_{xy} = \pm 1$ exactement lorsque $\mathbf{a}$ et $\mathbf{b}$ sont colinéaires, c'est-à-dire lorsque les points du nuage sont parfaitement alignés, et $r_{xy} = 0$ lorsque ces deux vecteurs sont orthogonaux : on dit alors que les variables sont **non corrélées**.

Remarque 5.2 **Diviser par n ou par $n-1$?**

Beaucoup de calculatrices, de manuels et de logiciels divisent par $n-1$ plutôt que par n dans la définition de la variance et de la covariance. Les deux formules ne répondent pas à la même question.
Si les n individus sont *tous* ceux qui nous intéressent, et que l'on veut seulement décrire leur dispersion, on divise par n : la variance est la moyenne des carrés des écarts. C'est notre situation, puisque nous cherchons à décrire la forme d'un nuage de points donné.
Si les n individus sont un *échantillon* tiré d'une population plus grande, dont on veut estimer la variance, on divise habituellement par $n-1$, du moins dans le cadre le plus courant d'un échantillon aléatoire simple. La raison est que l'on mesure les écarts à partir de la moyenne $\bar{x}$ de l'échantillon, faute de connaître celle de la population. Or $\bar{x}$ est justement le nombre c qui rend la somme $\sum(x_i - c)^2$ la plus petite possible : les écarts à $\bar{x}$ sont donc, dans l'ensemble, un peu plus petits que les écarts à la vraie moyenne, et la division par n sous-estime la variance de la population. On montre que, si les observations sont indépendantes et de même loi, elle la sous-estime en moyenne exactement par le facteur $\frac{n-1}{n}$, ce que la division par $n-1$ corrige. D'autres plans d'échantillonnage demandent d'autres ajustements : lorsqu'on tire sans remise dans une population finie, par exemple, on applique parfois une correction pour population finie. On peut aussi remarquer que les n écarts ont une somme nulle : dès que l'on en connaît $n-1$, le dernier est déterminé. Seuls $n-1$ écarts peuvent donc varier librement ; on dit que la somme des carrés compte $n-1$ degrés de liberté.
Pour les notes de mathématiques des six élèves, on obtient $\frac{300}{6} = 50$ en divisant par n, et $\frac{300}{5} = 60$ en divisant par $n-1$. L'écart entre les deux formules est notable pour un petit tableau, et négligeable lorsque n est grand. Il faut donc savoir ce que calcule l'outil que l'on utilise. Sur la plupart des calculatrices, la touche σ_x divise par n et la touche s_x par $n-1$; notre s_x correspond donc à la touche σ_x. En Python, `np.var` et `np.std` divisent par n par défaut, tandis que `np.cov` et les méthodes `var` et `std` de `pandas` divisent par $n-1$.
Pour l'ACP, ce choix n'a aucune importance : toutes les variances et covariances sont multipliées par le même facteur $\frac{n}{n-1}$, ce qui multiplie les valeurs propres par ce facteur sans changer ni les vecteurs propres ni les proportions de variance que nous calculerons, et les coefficients de corrélation ne changent pas du tout. Nous diviserons toujours par n, qui rend les formules plus simples.

Pour la classe 5.1 **La statistique à deux variables au secondaire**

Au deuxième cycle du secondaire, les élèves étudient les distributions à deux caractères : ils construisent des nuages de points, décrivent et estiment la corrélation linéaire, puis tracent une droite de régression, notamment par la méthode de Mayer ou par la méthode médiane-médiane [Min16]. Plusieurs idées de ce chapitre y sont déjà en germe. Le point moyen apparaît dans la méthode de Mayer, qui fait passer la droite par les points moyens de deux sous-groupes. Le coefficient de corrélation y est souvent estimé à partir de la forme du nuage, par exemple en comparant la longueur et la largeur du rectangle ou de l'ellipse qui l'entoure : un nuage très allongé indique une forte corrélation. L'ACP donne un sens précis à ces deux dimensions, la « longueur » et la « largeur » du nuage, qui deviendront les racines carrées de deux valeurs propres.

5.2 La droite qui représente le mieux un nuage de points

5.2.1 Deux façons de mesurer l'écart à une droite

Le nuage de la figure 5.1 s'allonge selon une direction oblique. Quelle est la droite qui le représente le mieux ? La question semble familière : c'est celle que l'on pose au secondaire lorsqu'on trace une droite de régression. Il faut pourtant préciser ce que l'on entend par « représenter le mieux ».

La **droite de régression** des moindres carrés est la droite $y = ax + b$ qui minimise la somme des carrés des écarts *verticaux* entre les points et la droite. Elle répond à une question de *prédiction* : connaissant la note de mathématiques x d'un élève, quelle note de sciences y peut-on prévoir ? On montre que cette droite passe par le point moyen et qu'elle a pour pente $a = s_{xy}/s_x^2$ (exercice 5.7). Ici, $a = 40/50 = 0{,}8$, et la droite de régression a pour équation $y = 75 + 0{,}8\,(x - 70)$, c'est-à-dire $y = 0{,}8\,x + 19$.

Mais les deux variables ne jouent pas le même rôle dans ce calcul. Si l'on cherchait plutôt à prédire la note de mathématiques à partir de celle de sciences, on minimiserait les écarts *horizontaux*, et l'on obtiendrait la droite $x = 70 + 0{,}8\,(y - 75)$, soit $y = 1{,}25\,x - 12{,}5$: une autre droite ! Lorsqu'aucune des deux variables n'a vocation à expliquer l'autre, et que l'on cherche simplement à décrire la forme du nuage, il est plus naturel de mesurer la distance de chaque point à la droite *perpendiculairement* à celle-ci (figure 5.2). C'est ce que nous ferons.

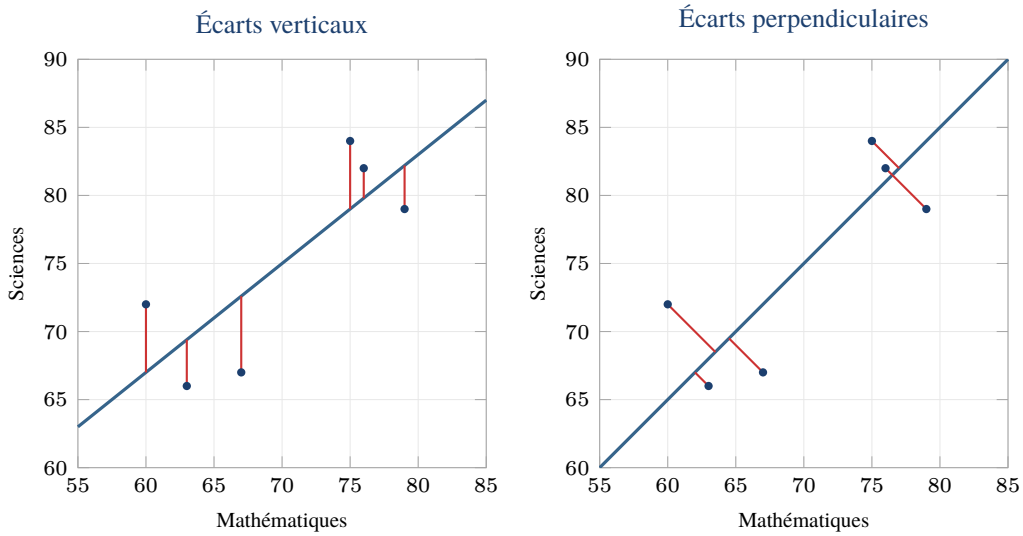


FIGURE 5.2 : À gauche, la droite de régression de y selon x minimise la somme des carrés des écarts verticaux (en rouge). À droite, la droite que nous cherchons minimise la somme des carrés des distances perpendiculaires. Ce n'est pas la même droite.

Définition 5.3 **Droite d'ajustement orthogonal**

Soit $\mathbf{x}_1, \dots, \mathbf{x}_n$ des points du plan (ou de $\mathbb{R}^p$), et $\mathcal{D}$ une droite. Notons d_i la distance du point $\mathbf{x}_i$ à la droite $\mathcal{D}$, c'est-à-dire la distance de $\mathbf{x}_i$ à sa projection orthogonale sur $\mathcal{D}$. La droite qui rend la somme

$$S(\mathcal{D}) = d_1^2 + d_2^2 + \cdots + d_n^2$$

la plus petite possible s'appelle la **droite d'ajustement orthogonal** du nuage. Nous verrons qu'elle porte aussi le nom d'**axe principal**.

5.2.2 On peut se ramener à des données centrées

Toutes les droites de régression passent par le point moyen ; c'est aussi le cas de la droite d'ajustement orthogonal. L'intuition est claire : si une droite passait à côté du centre de gravité du nuage, on pourrait la déplacer parallèlement à elle-même pour la rapprocher de l'ensemble des points. Nous admettrons ce résultat. Nous pouvons donc *centrer* les données, c'est-à-dire placer l'origine au point moyen, et ne chercher que parmi les droites qui passent par l'origine.

Dans toute la suite de cette section, les points $\mathbf{x}_1, \dots, \mathbf{x}_n$ *sont centrés* : ce sont les écarts à la moyenne des deux dernières colonnes du tableau 5.1, par exemple $\mathbf{x}_A = (-10, -3)$. Une droite passant par l'origine est l'espace engendré $\operatorname{span}\{\mathbf{u}\}$ par un vecteur unitaire $\mathbf{u}$.

5.2.3 Projeter sur une droite : le théorème de Pythagore

D'après la définition D.4, la projection orthogonale du point $\mathbf{x}_i$ sur la droite engendrée par le vecteur unitaire $\mathbf{u}$ est

$$\operatorname{proj}_{\mathbf{u}} \mathbf{x}_i = (\mathbf{x}_i \cdot \mathbf{u})\,\mathbf{u} = t_i\,\mathbf{u}, \qquad \text{où} \qquad t_i = \mathbf{x}_i \cdot \mathbf{u}.$$

Le nombre t_i est la *coordonnée* du point projeté sur la droite, graduée par $\mathbf{u}$: c'est la position, le long de la droite, de l'« ombre » du point $\mathbf{x}_i$. Le point $\mathbf{x}_i$, sa projection $t_i\mathbf{u}$ et l'origine forment un triangle rectangle (figure 5.3), et le théorème de Pythagore donne

$$\|\mathbf{x}_i\|^2 = t_i^2 + d_i^2. \tag{5.1}$$

Dans ce triangle, l'hypoténuse joint l'origine au point $\mathbf{x}_i$: sa longueur $\|\mathbf{x}_i\|$ est la distance du point à l'origine, c'est-à-dire au point moyen du nuage. Ni le point ni l'origine ne bougent lorsqu'on change de droite ; seule la droite pivote autour de l'origine. L'hypoténuse reste donc la même pour toutes les droites, et seuls les deux autres côtés, t_i et d_i, dépendent de $\mathbf{u}$.

C'est l'observation clé de tout le chapitre. En additionnant l'égalité (5.1) sur tous les points du nuage, on obtient

$$\underbrace{\sum_{i=1}^{n} \|\mathbf{x}_i\|^2}_{\text{ne dépend pas de } \mathbf{u}} = \sum_{i=1}^{n} t_i^2 + \sum_{i=1}^{n} d_i^2. \tag{5.2}$$

Le membre de gauche est la somme des carrés des distances des points à l'origine : il ne dépend que du nuage, pas de la droite choisie. Rendre la somme des d_i^2 la plus petite possible revient donc à rendre la somme des t_i^2 la plus grande possible. Or, les t_i ont une moyenne nulle, puisque $\frac{1}{n}\sum t_i = \left(\frac{1}{n}\sum \mathbf{x}_i\right) \cdot \mathbf{u} = \mathbf{0} \cdot \mathbf{u} = 0$ pour des données centrées ; la quantité $\frac{1}{n}\sum t_i^2$ est donc la *variance* des projections.

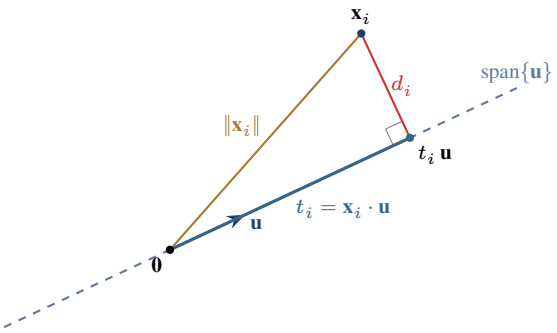


FIGURE 5.3 : La projection du point $\mathbf{x}_i$ sur la droite engendrée par le vecteur unitaire $\mathbf{u}$ est le point $t_i\mathbf{u}$, avec $t_i = \mathbf{x}_i \cdot \mathbf{u}$. Lorsque la droite pivote autour de l'origine, le point $\mathbf{x}_i$ et l'origine restent fixes : l'hypoténuse $\|\mathbf{x}_i\|$ ne change pas. Si le côté t_i s'allonge, le côté d_i raccourcit.

Proposition 5.1 **Deux problèmes équivalents**

Pour un nuage centré, la droite span$\{\mathbf{u}\}$ qui minimise la somme des carrés des distances des points à la droite est aussi celle qui maximise la variance des projections des points sur la droite.

Ces deux points de vue sont complémentaires. Le premier dit que la droite cherchée *passe au plus près des points* ; le second, que les points projetés sur cette droite sont *le plus étalés possible*, autrement dit que la projection écrase le moins possible le nuage. Pour reprendre l'image de l'introduction, c'est la direction selon laquelle l'ombre du nuage est la plus longue.

Exemple 5.1 **Trois directions pour les six élèves**

Comparons trois droites passant par le point moyen, de vecteurs directeurs unitaires $\mathbf{e}_1 = (1,0)$, $\mathbf{u} = \frac{1}{\sqrt{2}}(1,1)$ et $\mathbf{w} = \frac{1}{\sqrt{2}}(1,-1)$.

— Pour $\mathbf{e}_1$ (la droite horizontale), t_i est simplement l'écart $x_i - \bar{x}$, et la variance des projections est $s_x^2 = 50$.

— Pour $\mathbf{u}$, on obtient $t_i = \frac{1}{\sqrt{2}}((x_i - \bar{x}) + (y_i - \bar{y}))$, c'est-à-dire, pour les élèves A à F,

$$\sqrt{2}\,(t_A, t_B, \dots, t_F) = (-13, -16, -11, 14, 13, 13),$$

d'où

$$\frac{1}{6}\sum t_i^2 = \frac{169 + 256 + 121 + 196 + 169 + 169}{6 \cdot 2} = \frac{1080}{12} = 90.$$

— Pour $\mathbf{w}$, on trouve de même $\sqrt{2}\,(t_A, \dots, t_F) = (-7, 2, 5, -4, -1, 5)$ et une variance de $\frac{120}{12} = 10$.

Enfin, $\frac{1}{6}\sum\|\mathbf{x}_i\|^2 = s_x^2 + s_y^2 = 100$. Pour la direction $\mathbf{u}$, la relation (5.2), divisée par n, s'écrit donc $100 = 90 + 10$: la variance des projections vaut 90, et la moyenne des carrés des distances à la droite vaut 10. Pour $\mathbf{w}$, c'est l'inverse, $100 = 10 + 90$. La direction $\mathbf{u}$ est de loin la meilleure des trois.

5.2.4 La variance dans une direction quelconque

Plutôt que d'essayer les directions une à une, exprimons la variance des projections en fonction de $\mathbf{u}$. Si $\mathbf{x}_i = (a_i, b_i)$ et $\mathbf{u} = (u_1, u_2)$, alors $t_i = a_i u_1 + b_i u_2$, et

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} t_i^2 &= \frac{1}{n}\sum_{i=1}^{n}(a_i^2 u_1^2 + 2a_i b_i u_1 u_2 + b_i^2 u_2^2)\\ &= s_x^2\, u_1^2 + 2s_{xy}\, u_1 u_2 + s_y^2\, u_2^2.\end{aligned}$$

On reconnaît une expression déjà rencontrée : si l'on pose

$$C = \begin{pmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{pmatrix},$$

alors $C\mathbf{u} = (s_x^2 u_1 + s_{xy} u_2,\ s_{xy} u_1 + s_y^2 u_2)$ et

$$\frac{1}{n}\sum_{i=1}^{n} t_i^2 = \mathbf{u} \cdot C\mathbf{u} = \mathbf{u}^{\mathsf{T}} C\, \mathbf{u}. \tag{5.3}$$

La matrice C s'appelle la **matrice de covariance** des deux variables. Elle contient les deux variances sur sa diagonale et la covariance ailleurs, et elle est *symétrique*. Pour nos six élèves,

$$C = \begin{pmatrix} 50 & 40 \\ 40 & 50 \end{pmatrix}.$$

Toute direction du plan s'écrit $\mathbf{u} = (\cos\theta, \sin\theta)$, où θ est l'angle que fait la droite avec l'axe horizontal. La variance des projections devient alors une fonction de θ :

$$f(\theta) = 50\cos^2\theta + 80\cos\theta\sin\theta + 50\sin^2\theta = 50 + 40\sin(2\theta),$$

puisque $\cos^2\theta + \sin^2\theta = 1$ et $2\sin\theta\cos\theta = \sin(2\theta)$. Cette fonction sinusoïdale atteint son maximum, 90, lorsque $2\theta = 90°$, donc pour $\theta = 45°$, et son minimum, 10, pour $\theta = 135°$ (figure 5.4). La droite d'ajustement orthogonal est donc la droite de pente 1 qui passe par le point moyen $(70, 75)$, c'est-à-dire la droite d'équation $y = x + 5$.

Pour la classe 5.2 — **Une fonction sinusoïdale**

> La fonction $f(\theta) = 50 + 40\sin(2\theta)$ est une fonction sinusoïdale, du type de celles qu'étudient les élèves de cinquième secondaire (séquence SN), avec ses paramètres habituels : une valeur moyenne de 50, une amplitude de 40 et une période de 180°. La période s'explique sans calcul : tourner une droite d'un demi-tour la ramène sur elle-même. Le problème d'optimisation de cette section est donc accessible à un élève du secondaire, pourvu qu'on lui fournisse l'expression de f ; le chercher par essais, avec une droite que l'on fait tourner autour du point moyen dans GeoGebra, est aussi une activité éclairante.

5.2.5 Le lien avec les vecteurs propres

Le maximum est atteint dans la direction de $\mathbf{u} = \frac{1}{\sqrt{2}}(1, 1)$. Ce vecteur n'est pas quelconque : la matrice $C = \begin{pmatrix} 50 & 40 \\ 40 & 50 \end{pmatrix}$ a pour valeurs propres $\lambda_1 = 90$ et $\lambda_2 = 10$, de vecteurs propres respectifs $(1, 1)$ et $(1, -1)$. Le calcul est exactement celui de l'exemple D.7, puisque $C = 10\begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}$ et que, comme $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$, cette matrice a des coefficients diagonaux

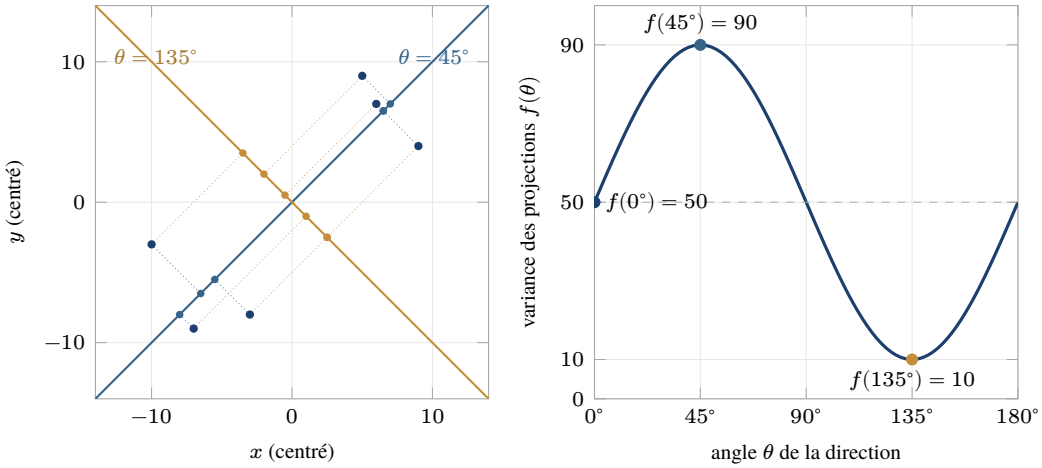


FIGURE 5.4 : À gauche, le nuage centré et les projections de ses points sur deux droites : celle d'angle 45° (en bleu), où les projections sont très étalées, et celle d'angle 135° (en ocre), où elles sont tassées près de l'origine. À droite, la variance $f(\theta) = 50 + 40\sin(2\theta)$ des projections, en fonction de l'angle de la droite.

égaux. La droite d'ajustement orthogonal est portée par le vecteur propre associé à la *plus grande* valeur propre ; cette plus grande valeur propre est la variance maximale, et la seconde valeur propre est la variance minimale.

Ce n'est pas une coïncidence. Pour une matrice symétrique C quelconque, posons $\mathbf{u}(\theta) = (\cos\theta, \sin\theta)$ et $f(\theta) = \mathbf{u}^\intercal C\,\mathbf{u}$. La dérivée de $\mathbf{u}(\theta)$ est le vecteur $\mathbf{u}'(\theta) = (-\sin\theta, \cos\theta)$, qui est unitaire et perpendiculaire à $\mathbf{u}$. Par la règle de dérivation d'un produit (qui s'applique au produit scalaire composante par composante),

$$f'(\theta) = \mathbf{u}'^\intercal C\,\mathbf{u} + \mathbf{u}^\intercal C\,\mathbf{u}' = 2\,\mathbf{u}' \cdot C\mathbf{u},$$

où l'on a utilisé la symétrie de C : $\mathbf{u}^\intercal C\,\mathbf{u}' = (\mathbf{u}^\intercal C\,\mathbf{u}')^\intercal = \mathbf{u}'^\intercal C^\intercal \mathbf{u} = \mathbf{u}'^\intercal C\,\mathbf{u}$. La dérivée s'annule donc exactement lorsque $C\mathbf{u}$ est perpendiculaire à $\mathbf{u}'$, c'est-à-dire, dans le plan, lorsque $C\mathbf{u}$ est *parallèle* à $\mathbf{u}$: lorsque $C\mathbf{u} = \lambda\mathbf{u}$ pour un certain nombre λ. Les directions où la variance est maximale ou minimale sont donc des directions propres de C ! De plus, si $C\mathbf{u} = \lambda\mathbf{u}$ avec $\mathbf{u}$ unitaire, alors $f = \mathbf{u} \cdot \lambda\mathbf{u} = \lambda$: la variance dans une direction propre est la valeur propre correspondante.

Théorème 5.2 **La droite d'ajustement orthogonal**

Soit un nuage de points du plan, de point moyen M et de matrice de covariance C, et soit $\lambda_1 \geq \lambda_2$ les valeurs propres de C, de vecteurs propres unitaires $\mathbf{v}_1$ et $\mathbf{v}_2$.

1. La droite d'ajustement orthogonal est la droite qui passe par M et qui est dirigée par $\mathbf{v}_1$.
2. La variance des projections des points sur cette droite vaut λ_1, et la moyenne des carrés des distances des points à cette droite vaut λ_2.
3. La variance totale du nuage, $s_x^2 + s_y^2$, est égale à $\lambda_1 + \lambda_2$.

Démonstration. Le point (1) découle du fait, admis plus haut, que la droite d'ajustement orthogonal passe par le point moyen, de la proposition 5.1 et de l'étude de f qui précède : f est continue et périodique, donc elle atteint son maximum, en un point où $f' = 0$,

c'est-à-dire dans une direction propre ; parmi les deux valeurs propres, la plus grande est λ_1. Le point (3) est la propriété de la trace (section D.10) : la somme des valeurs propres est égale à la somme $s_x^2 + s_y^2$ des coefficients diagonaux. Le point (2) en découle, en divisant l'égalité (5.2) par n : la moyenne des d_i^2 vaut $(s_x^2 + s_y^2) - \lambda_1 = \lambda_2$. □

Le théorème donne aussi un sens précis à la « forme » du nuage. La variance totale $s_x^2 + s_y^2 = 100$ des six élèves se répartit en $\lambda_1 = 90$ le long de l'axe principal et $\lambda_2 = 10$ dans la direction perpendiculaire. On dit que l'axe principal **explique** $\frac{90}{100} = 90\,\%$ de la variance totale : en remplaçant chaque point par sa projection sur la droite, on ne perd que $10\,\%$ de la dispersion du nuage. Géométriquement, le nuage ressemble à une ellipse dont les axes sont dirigés par les vecteurs propres, et dont les demi-axes sont proportionnels à $\sqrt{\lambda_1}$ et à $\sqrt{\lambda_2}$ (figure 5.5) ; on retrouve l'ellipse de la figure D.10.

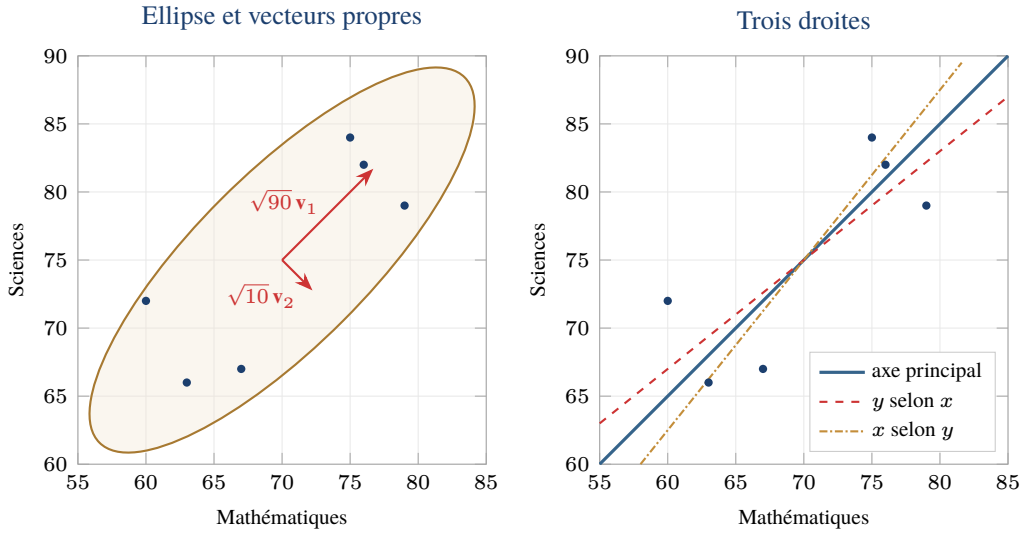


FIGURE 5.5 : À gauche, les vecteurs propres de la matrice de covariance, multipliés par la racine carrée de leur valeur propre (l'écart type dans chaque direction), dessinent les axes d'une ellipse qui épouse la forme du nuage ; on a tracé l'ellipse de demi-axes $2\sqrt{90}$ et $2\sqrt{10}$. À droite, l'axe principal $y = x + 5$ se situe entre les deux droites de régression, $y = 0{,}8\,x + 19$ et $y = 1{,}25\,x - 12{,}5$.

Les coordonnées t_i des projections sur l'axe principal ont, elles aussi, une interprétation. Puisque $t_i = \frac{1}{\sqrt{2}}\big((x_i - 70) + (y_i - 75)\big)$, elles sont proportionnelles à la *somme* des écarts à la moyenne dans les deux matières : la coordonnée sur l'axe principal mesure le niveau général de l'élève. La coordonnée sur le second axe, $\frac{1}{\sqrt{2}}\big((x_i - 70) - (y_i - 75)\big)$, mesure plutôt la *différence* entre les deux matières : elle est positive pour un élève relativement plus fort en mathématiques qu'en sciences. Ce passage d'un couple de notes à un couple « niveau général, préférence » est un simple changement de base orthonormée, une rotation de $45°$ du repère : il ne fait perdre aucune information, mais il range l'information par ordre d'importance.

Remarque 5.3 **Régression ou axe principal ?**

La figure 5.5 illustre un fait général : l'axe principal se situe toujours entre les deux droites de régression, et les trois droites sont confondues seulement lorsque les points sont parfaitement alignés.
Pourquoi deux droites de régression ? Parce qu'elles répondent à deux questions différentes. La droite de régression de y selon x sert à prédire y lorsque x est connu : seule l'erreur commise sur y compte, d'où les écarts verticaux. La droite de régression de x selon y sert à prédire x lorsque y est connu ; elle minimise les écarts horizontaux. Seul l'axe principal traite les deux variables de façon symétrique.
Les droites de régression sont « tirées » vers l'horizontale ou vers la verticale. Découpons le nuage en tranches verticales : la droite de régression de y selon x passe, à peu près, par la moyenne des y de chaque tranche. Dans une tranche, les points s'étalent de part et d'autre de l'axe principal et, comme la corrélation n'est pas parfaite, leur moyenne est plus proche de la moyenne générale $\bar{y}$ que ne l'indique l'axe. Pour les six élèves, un élève qui a 80 en mathématiques, soit 10 points au-dessus de la moyenne, reçoit une note prédite en sciences de $75 + 0{,}8 \times 10 = 83$, seulement 8 points au-dessus de la moyenne, alors que l'axe principal indiquerait 85. Si l'on prédit ensuite la note de mathématiques à partir de 83 en sciences, avec l'autre droite, on obtient $70 + 0{,}8 \times 8 = 76{,}4$, et non 80 : chaque prédiction rapproche de la moyenne. En unités d'écart type, un écart de k écarts types sur x fait prédire un écart de r_{xy} k écarts types sur y ; comme $|r_{xy}| < 1$, la prédiction est toujours moins extrême que la donnée de départ.
Francis Galton a observé ce phénomène en étudiant la taille des enfants en fonction de celle de leurs parents [Gal86] : les enfants de parents très grands sont grands, mais en moyenne moins grands que leurs parents. Il l'a appelé la *régression vers la moyenne*, d'où le nom de la méthode. Il ne s'agit pas d'un resserrement des tailles d'une génération à l'autre, puisque l'effet joue aussi dans l'autre sens : les parents d'enfants très grands sont, en moyenne, moins grands que leurs enfants. Ce sont exactement les deux droites de régression.
Le choix entre les deux méthodes dépend de la question posée : pour *prédire* une variable à partir d'une autre, on utilise la régression ; pour *décrire* un nuage sans privilégier de variable, l'axe principal. Il faut toutefois retenir une faiblesse de l'axe principal : il dépend des unités de mesure (exercice 5.4), ce qui n'est pas le cas de la droite de régression, ni du coefficient de corrélation.

5.3 Des données dans l'espace

5.3.1 Les manchots de l'archipel Palmer

Entre 2007 et 2009, l'écologiste Kristen Gorman et ses collègues de la station de recherche Palmer, en Antarctique, ont mesuré plusieurs centaines de manchots appartenant à trois espèces : le manchot Adélie, le manchot à jugulaire et le manchot papou [GWF14]. Ces données, rendues publiques sous une licence libre [HHG20], sont aujourd'hui très utilisées pour enseigner la statistique et la science des données. Nous retiendrons, pour $n = 342$ manchots, trois variables mesurées en millimètres : la longueur du bec, la hauteur du bec (son épaisseur, mesurée verticalement à la base) et la longueur de la nageoire (tableau 5.2). Chaque manchot est donc un point de $\mathbb{R}^3$.

Un nuage de points de l'espace se dessine en perspective, mais une perspective n'est rien

Table 5.2 : Trois manchots de l'archipel Palmer, un de chaque espèce, puis les moyennes et les écarts types des trois variables sur l'ensemble des 342 manchots.

	Espèce	Longueur du bec	Hauteur du bec	Nageoire
manchot 1	Adélie	39,1	18,7	181
manchot 153	papou	46,1	13,2	211
manchot 277	à jugulaire	46,5	17,9	192
moyenne		43,92	17,15	200,92
écart type		5,45	1,97	14,04

d'autre qu'une projection sur le plan de la page, et la figure 5.6 n'en montre qu'un point de vue parmi d'autres. Sur un écran, on peut faire tourner le nuage avec la souris, ce qui revient à faire défiler un grand nombre de projections ; sur papier, il faut choisir.

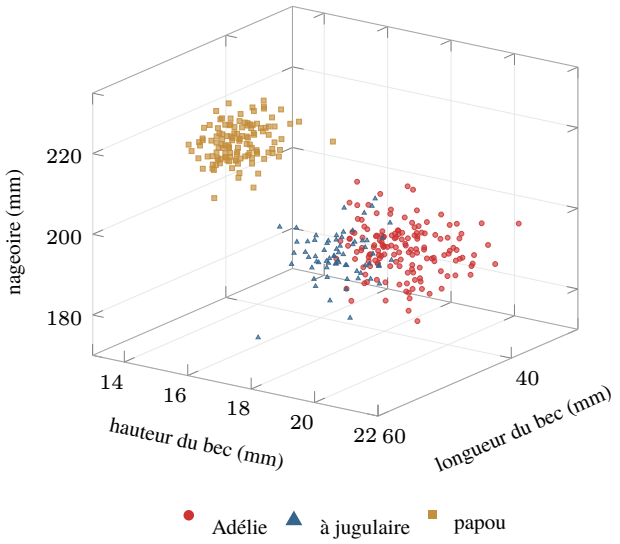


Figure 5.6 : Les 342 manchots vus en perspective. Les couleurs et les formes des marques indiquent l'espèce. Cette vue suggère trois groupes, mais on ne peut pas y lire les trois mesures d'un manchot : le dessin n'a que deux dimensions, et deux points de l'espace situés l'un derrière l'autre, dans la direction du regard, sont dessinés au même endroit de la page.

5.3.2 Regarder le nuage sous différents angles

La façon la plus simple de représenter des données de dimension 3 sur une page consiste à oublier une des variables, c'est-à-dire à projeter le nuage sur l'un des trois plans de coordonnées. La figure 5.7 montre les trois projections obtenues. Elles ne racontent pas toutes la même histoire.

Dans le premier plan (longueur et hauteur du bec), les trois espèces sont bien séparées : les manchots Adélie ont un bec court et haut, les papous un bec long et mince, et les manchots à jugulaire un bec à la fois long et haut. Dans le troisième plan (hauteur du bec et nageoire), les papous, qui ont de longues nageoires et un bec mince, se détachent clairement, mais les deux autres espèces se superposent presque entièrement : en ne regardant que ce plan, on ne soupçonnerait pas qu'il y a là deux espèces. Le choix du plan de projection n'est donc pas anodin. Rien n'oblige d'ailleurs à se limiter aux plans de coordonnées : on peut projeter le nuage sur n'importe quel plan passant par le point moyen. Une question naturelle se pose alors : *quel est le meilleur plan sur lequel projeter*

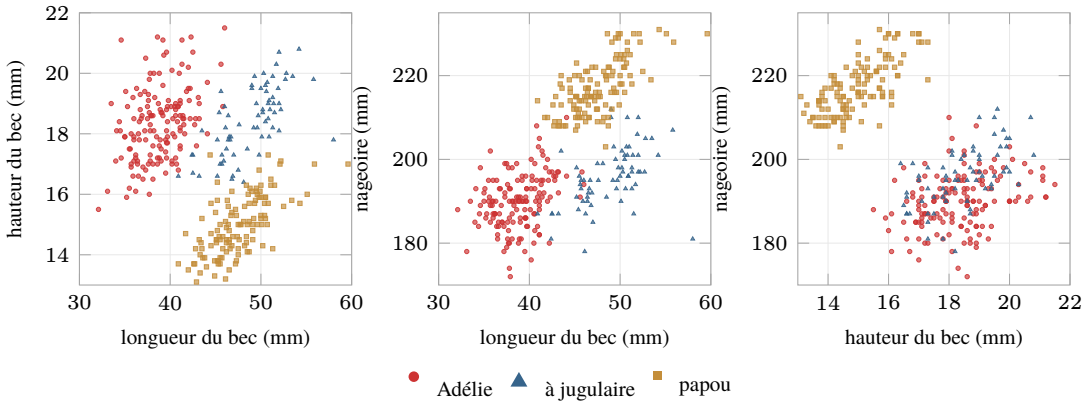


FIGURE 5.7 : Les projections du nuage des manchots sur les trois plans de coordonnées. Dans le premier, les trois espèces forment trois groupes bien distincts. Dans le troisième, les manchots papous se détachent nettement, mais les manchots Adélie et à jugulaire se confondent.

les données ?

Il faut d'abord s'entendre sur le sens de « meilleur ». Nous ne pouvons pas demander à la méthode de séparer les espèces, puisqu'elle ne les connaît pas : les couleurs ont été ajoutées après coup, et dans la plupart des applications, on ne dispose d'aucune étiquette de ce genre. Nous reprenons plutôt le critère de la section précédente : le meilleur plan est celui qui passe au plus près des points, ou, ce qui revient au même, celui sur lequel le nuage projeté est le plus étalé.

5.3.3 Projeter sur un plan

Soit Π un plan passant par l'origine, et $(\mathbf{a}, \mathbf{b})$ une base orthonormée de ce plan. La projection orthogonale d'un point $\mathbf{x}$ sur Π est

$$\operatorname{proj}_\Pi \mathbf{x} = (\mathbf{x} \cdot \mathbf{a})\,\mathbf{a} + (\mathbf{x} \cdot \mathbf{b})\,\mathbf{b} :$$

on projette sur chacune des deux droites de la base, et l'on additionne. Le vecteur $\mathbf{x} - \operatorname{proj}_\Pi \mathbf{x}$ est orthogonal à $\mathbf{a}$, puisque

$$(\mathbf{x} - \operatorname{proj}_\Pi \mathbf{x}) \cdot \mathbf{a} = \mathbf{x} \cdot \mathbf{a} - (\mathbf{x} \cdot \mathbf{a})(\mathbf{a} \cdot \mathbf{a}) - (\mathbf{x} \cdot \mathbf{b})(\mathbf{b} \cdot \mathbf{a}) = \mathbf{x} \cdot \mathbf{a} - \mathbf{x} \cdot \mathbf{a} - 0 = 0,$$

et de même à $\mathbf{b}$; il est donc orthogonal au plan tout entier ; sa norme est la distance d du point $\mathbf{x}$ au plan. Les nombres $\mathbf{x} \cdot \mathbf{a}$ et $\mathbf{x} \cdot \mathbf{b}$ sont les coordonnées du point projeté dans la base $(\mathbf{a}, \mathbf{b})$: ce sont eux que l'on porte en abscisse et en ordonnée pour dessiner le nuage projeté. Le théorème de Pythagore s'applique encore :

$$\|\mathbf{x}\|^2 = (\mathbf{x} \cdot \mathbf{a})^2 + (\mathbf{x} \cdot \mathbf{b})^2 + d^2.$$

Le même raisonnement qu'à la section 5.2 montre donc que, pour un nuage centré, le plan qui minimise la somme des carrés des distances des points au plan est celui qui maximise

$$\frac{1}{n} \sum_{i=1}^{n} \Big((\mathbf{x}_i \cdot \mathbf{a})^2 + (\mathbf{x}_i \cdot \mathbf{b})^2 \Big) = \mathbf{a}^\mathsf{T} C\, \mathbf{a} + \mathbf{b}^\mathsf{T} C\, \mathbf{b},$$

la variance du nuage projeté, où C est maintenant la matrice de covariance 3×3 des trois variables. Ses coefficients diagonaux sont les trois variances, et le coefficient (j, k) est la covariance des variables j et k ; l'égalité découle de (5.3), appliquée à $\mathbf{a}$ puis à $\mathbf{b}$ (nous la justifierons en toute dimension à la section 5.4).

5.3.4 Un problème d'unités : réduire les variables

Avant de chercher ce plan, un obstacle doit être levé. La longueur de la nageoire a un écart type de 14,04 mm, la hauteur du bec un écart type de 1,97 mm seulement. Une variable très dispersée pèse beaucoup plus lourd dans la variance totale : à elle seule, la nageoire contribue $14{,}04^2 \approx 197$ des quelque 231 mm^2 de variance totale du nuage. Le meilleur plan serait donc presque imposé par la nageoire, simplement parce qu'elle est plus grande que le bec, et non parce qu'elle est plus informative. Le problème serait pire encore avec des variables mesurées dans des unités différentes, comme une longueur en millimètres et une masse en grammes : le résultat dépendrait alors du choix arbitraire des unités (exercice 5.4).

La solution habituelle consiste à **réduire** les variables, en divisant chaque écart à la moyenne par l'écart type :

$$z_i = \frac{x_i - \bar{x}}{s_x}.$$

Le nombre z_i, appelé la **cote z** de la valeur x_i, est sans unité : il indique de combien d'écarts types la valeur x_i s'éloigne de la moyenne. Une variable **centrée réduite** a une moyenne nulle et une variance égale à 1, et la covariance de deux variables centrées réduites est égale à leur coefficient de corrélation. La matrice de covariance des variables réduites est donc la **matrice de corrélation** R des variables de départ, dont tous les coefficients diagonaux valent 1. Pour les manchots,

$$R = \begin{pmatrix} 1 & -0{,}235 & 0{,}656 \\ -0{,}235 & 1 & -0{,}584 \\ 0{,}656 & -0{,}584 & 1 \end{pmatrix},$$

où les variables sont, dans l'ordre, la longueur du bec, la hauteur du bec et la longueur de la nageoire. On y lit, par exemple, que les manchots qui ont une longue nageoire ont tendance à avoir un bec long ($r = 0{,}656$) et mince ($r = -0{,}584$).

Pour la classe 5.3 — **La cote z et la cote R**

> Les futurs enseignants québécois connaissent bien un usage de la cote z : la *cote de rendement au collégial* (cote R), qui sert à l'admission dans les programmes universitaires contingentés, est construite à partir de la cote z d'un étudiant dans chacun de ses groupes-cours. L'idée est exactement celle de la réduction : une note de 75 n'a pas la même valeur dans un groupe où la moyenne est de 65 que dans un groupe où elle est de 85, et l'écart à la moyenne, exprimé en écarts types, permet de comparer des résultats obtenus dans des contextes différents. En classe, la cote z permet de répondre à une question que se posent souvent les élèves : « ai-je mieux réussi en mathématiques ou en français ? »

5.3.5 Le meilleur plan

Comme dans le plan, on montre que la meilleure droite est dirigée par le vecteur propre de R associé à la plus grande valeur propre. Pour le meilleur plan, la réponse est tout aussi simple : *il est engendré par les vecteurs propres associés aux deux plus grandes valeurs propres* (théorème 5.4). La matrice R étant symétrique, le théorème spectral (théorème D.4) garantit que ces vecteurs propres sont orthogonaux. Un calcul numérique (section 5.4.4) donne les valeurs propres

$$\lambda_1 = 2{,}003, \qquad \lambda_2 = 0{,}767, \qquad \lambda_3 = 0{,}230,$$

dont la somme vaut 3, la trace de R, et les vecteurs propres unitaires

$$\mathbf{v}_1 = \begin{pmatrix} 0{,}551 \\ -0{,}512 \\ 0{,}659 \end{pmatrix}, \qquad \mathbf{v}_2 = \begin{pmatrix} 0{,}657 \\ 0{,}753 \\ 0{,}036 \end{pmatrix}, \qquad \mathbf{v}_3 = \begin{pmatrix} -0{,}514 \\ 0{,}413 \\ 0{,}752 \end{pmatrix}.$$

Le plan engendré par $\mathbf{v}_1$ et $\mathbf{v}_2$ conserve $\frac{2{,}003+0{,}767}{3} \approx 92{,}3\,\%$ de la variance totale du nuage réduit : en projetant sur ce plan, on ne perd que $7{,}7\,\%$ de la dispersion. La figure 5.8 montre le résultat. Les trois espèces y apparaissent clairement, alors que la méthode ignorait leur existence. À titre de comparaison, le plan engendré par $\mathbf{v}_2$ et $\mathbf{v}_3$ ne conserve que $\frac{0{,}767+0{,}230}{3} \approx 33{,}2\,\%$ de la variance, et les manchots Adélie et papous y sont indiscernables.

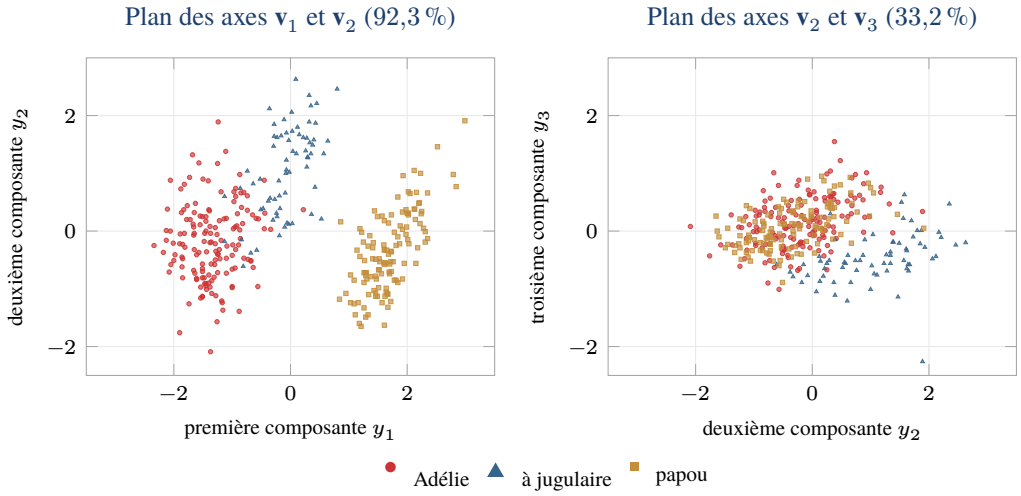


FIGURE 5.8 : À gauche, la projection du nuage réduit des manchots sur le plan engendré par les deux premiers vecteurs propres de la matrice de corrélation : c'est le meilleur plan, au sens de la variance. À droite, la projection sur le plan engendré par les deux derniers vecteurs propres, qui est le pire : le nuage y est écrasé.

Les composantes des vecteurs propres permettent d'interpréter les nouveaux axes. La coordonnée d'un manchot sur le premier axe,

$$y_1 = 0{,}551\, z_{\text{long. bec}} - 0{,}512\, z_{\text{haut. bec}} + 0{,}659\, z_{\text{nageoire}},$$

est grande pour un manchot à long bec mince et à longue nageoire : c'est le profil du manchot papou, que l'on trouve à droite de la figure. La coordonnée sur le second axe, $y_2 \approx 0{,}657\, z_{\text{long. bec}} + 0{,}753\, z_{\text{haut. bec}}$, ne dépend presque pas de la nageoire ; elle mesure la *taille du bec*, dans ses deux dimensions, et elle sépare les manchots à jugulaire, au bec long et haut, des manchots Adélie. L'ACP a ainsi remplacé trois mesures corrélées par deux nouvelles variables, des combinaisons linéaires des premières, qui résument l'essentiel de l'information.

La version web de ces notes contient une application interactive qui permet de faire cette recherche soi-même : en faisant tourner le nuage, on choisit le plan de l'écran, et l'on voit l'ombre du nuage sur ce plan, avec la part de la variance qu'il conserve.

Outil 5.1 **Faire tourner un nuage de points**

Pour développer l'intuition, rien ne vaut un nuage que l'on peut faire tourner. Les calculatrices 3D de GeoGebra et de Desmos tracent un nuage de points de l'espace à partir d'une liste de triplets : on peut y coller les mesures des manchots et chercher, en faisant tourner la figure, le point de vue où les trois espèces se séparent le mieux. En Python, la bibliothèque `plotly` produit un nuage 3D que l'on fait pivoter à la souris (`plotly.express.scatter_3d`). Pour des données de dimension supérieure, l'outil en ligne *Embedding Projector* de TensorFlow (`https://projector.tensorflow.org`) calcule l'ACP d'un jeu de données et en affiche la projection sur les trois premiers axes principaux, dans une vue que l'on peut faire tourner.

5.4 L'analyse en composantes principales

Nous pouvons maintenant décrire la méthode en toute généralité. Les deux exemples précédents en contiennent tous les ingrédients ; il ne reste qu'à les écrire pour p variables, avec des matrices.

5.4.1 Le tableau centré et la matrice de covariance

On dispose de n individus et de p variables. Après avoir centré chaque variable (et l'avoir réduite, s'il y a lieu : voir la remarque 5.4), on range les données dans une matrice X de taille $n \times p$, dont la ligne i est le vecteur $\mathbf{x}_i^\intercal$ de l'individu i. Puisque chaque colonne est centrée, la somme des lignes est le vecteur nul.

Définition 5.4 **Matrice de covariance**

La **matrice de covariance** du tableau centré X est la matrice $p \times p$

$$C = \frac{1}{n} X^\intercal X.$$

Son coefficient (j, k) est $c_{jk} = \frac{1}{n} \sum_{i=1}^{n} x_{ij} x_{ik}$, la covariance des variables j et k ; ses coefficients diagonaux sont les variances. Si les variables ont été réduites, C est la matrice de corrélation des variables de départ.

Proposition 5.3 **Propriétés de la matrice de covariance**

Soit C la matrice de covariance d'un tableau centré X.

1. C est symétrique.
2. Pour tout vecteur unitaire $\mathbf{u} \in \mathbb{R}^p$, le nombre $\mathbf{u}^\intercal C\,\mathbf{u}$ est la variance des projections $t_i = \mathbf{x}_i \cdot \mathbf{u}$ des individus sur la droite engendrée par $\mathbf{u}$. En particulier, $\mathbf{u}^\intercal C\,\mathbf{u} \geq 0$.
3. Les valeurs propres de C sont réelles et positives ou nulles, et leur somme est la **variance totale** $c_{11} + c_{22} + \cdots + c_{pp}$ du nuage, c'est-à-dire la moyenne des carrés des distances des individus au point moyen.

Démonstration. (1) $C^\intercal = \frac{1}{n} X^\intercal (X^\intercal)^\intercal = \frac{1}{n} X^\intercal X = C$.

(2) Le vecteur $X\mathbf{u}$ a pour composantes les nombres $t_i = \mathbf{x}_i \cdot \mathbf{u}$, dont la moyenne est nulle, et
$$\mathbf{u}^\mathsf{T} C\,\mathbf{u} = \frac{1}{n}\,\mathbf{u}^\mathsf{T} X^\mathsf{T} X\mathbf{u} = \frac{1}{n}\,(X\mathbf{u})\cdot(X\mathbf{u}) = \frac{1}{n}\sum_{i=1}^{n} t_i^2.$$

(3) D'après le théorème spectral, les valeurs propres de C sont réelles. Si $\mathbf{v}$ est un vecteur propre unitaire associé à λ, alors $\lambda = \mathbf{v}^\mathsf{T}(\lambda\mathbf{v}) = \mathbf{v}^\mathsf{T} C\,\mathbf{v} \geq 0$ par (2). La somme des valeurs propres est la trace de C : avec $C = QDQ^\mathsf{T}$ et $\operatorname{tr}(AB) = \operatorname{tr}(BA)$, on a $\operatorname{tr}(C) = \operatorname{tr}(DQ^\mathsf{T}Q) = \operatorname{tr}(D)$, puisque $Q^\mathsf{T}Q = I$. Enfin $\sum_j c_{jj} = \frac{1}{n}\sum_i \sum_j x_{ij}^2 = \frac{1}{n}\sum_i \|\mathbf{x}_i\|^2$.

□

Remarque 5.4 **Faut-il toujours réduire les variables ?**

Non. Réduire, c'est décider que toutes les variables ont la même importance au départ : chacune reçoit une variance de 1, quelle que soit sa dispersion réelle. Cette décision est généralement souhaitable dans deux cas :

— lorsque les variables s'expriment dans des unités différentes (des millimètres et des grammes, des secondes et des mètres), car la variance totale additionnerait alors des quantités qui ne se comparent pas, et le résultat dépendrait du choix des unités (exercices 5.4 et 5.9) ;

— lorsque les variables ont la même unité, mais des ordres de grandeur si différents que la plus dispersée imposerait à elle seule le premier axe, comme la nageoire et le bec des manchots.

En revanche, lorsque les variables s'expriment dans la même unité et que leurs dispersions se comparent, les différences de variance font partie de ce que l'on veut décrire, et la réduction les efface. Si les points sont des positions dans l'espace, par exemple, réduire revient à étirer chaque axe séparément, ce qui déforme le nuage dont on cherche justement la forme. De même, pour des notes sur 100 à plusieurs examens, un examen où tous les élèves ont obtenu à peu près la même note ne distingue presque pas les élèves ; une fois réduit, il pèserait pourtant autant que les autres, et ses petits écarts, dus en bonne partie au hasard, seraient amplifiés. Une variable constante ne peut d'ailleurs pas être réduite du tout, puisque son écart type est nul.

Les deux analyses ne donnent pas les mêmes axes : l'une diagonalise la matrice de covariance, l'autre la matrice de corrélation. Il faut donc toujours préciser laquelle on a effectuée. Dans le doute, et en particulier dès que les unités diffèrent, on réduit ; c'est ce que font par défaut plusieurs logiciels de statistique.

5.4.2 Axes principaux et composantes principales

Le théorème spectral (théorème D.4) permet d'écrire $C = QDQ^\mathsf{T}$, où D est la matrice diagonale des valeurs propres et où les colonnes de Q forment une base orthonormée de vecteurs propres. On numérote les valeurs propres par ordre décroissant, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, et l'on note $\mathbf{v}_1, \ldots, \mathbf{v}_p$ les vecteurs propres correspondants.

Définition 5.5 **Analyse en composantes principales**

Soit X un tableau centré, C sa matrice de covariance, et $(\mathbf{v}_1, \dots, \mathbf{v}_p)$ une base orthonormée de vecteurs propres de C, associés aux valeurs propres $\lambda_1 \geq \cdots \geq \lambda_p$.

1. La droite engendrée par $\mathbf{v}_k$ est le k^{e} **axe principal** du nuage.
2. La k^{e} **composante principale** est la variable y_k dont la valeur pour l'individu i est la coordonnée de sa projection sur le k^{e} axe :

$$y_{ik} = \mathbf{x}_i \cdot \mathbf{v}_k.$$

   Matriciellement, le vecteur des valeurs de y_k est $X\mathbf{v}_k$, et le tableau des composantes principales est $Y = XQ$.
3. La **proportion de variance expliquée** par le k^{e} axe est

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}.$$

Le vocabulaire varie d'un ouvrage à l'autre : certains appellent « composantes principales » les vecteurs $\mathbf{v}_k$ eux-mêmes, et les logiciels désignent souvent les valeurs y_{ik} par le mot anglais *scores*. L'essentiel est de distinguer les *directions* $\mathbf{v}_k$, qui sont des vecteurs de $\mathbb{R}^p$, des *coordonnées* y_{ik} des n individus dans ces directions.

Théorème 5.4 **Propriétés de l'ACP**

Avec les notations de la définition 5.5 :

1. chaque composante principale y_k est de moyenne nulle et de variance λ_k ;
2. deux composantes principales distinctes y_j et y_k sont non corrélées ;
3. pour tout $q \leq p$, parmi tous les sous-espaces de dimension q de $\mathbb{R}^p$, celui qui est engendré par $\mathbf{v}_1, \dots, \mathbf{v}_q$ est celui sur lequel la variance du nuage projeté est la plus grande ; cette variance maximale vaut $\lambda_1 + \cdots + \lambda_q$;
4. de façon équivalente, si l'on remplace chaque individu par sa projection sur ce sous-espace, la moyenne des carrés des distances entre les individus et leurs projections vaut $\lambda_{q+1} + \cdots + \lambda_p$, et aucun autre sous-espace de dimension q ne fait mieux.

Démonstration. (1) et (2). La moyenne de y_k est $\left(\frac{1}{n}\sum_i \mathbf{x}_i\right) \cdot \mathbf{v}_k = 0$. La covariance de y_j et de y_k vaut

$$\frac{1}{n}\sum_{i=1}^{n} y_{ij}\, y_{ik} = \frac{1}{n}\,(X\mathbf{v}_j) \cdot (X\mathbf{v}_k) = \mathbf{v}_j^{\mathsf{T}} C\, \mathbf{v}_k = \lambda_k\, \mathbf{v}_j \cdot \mathbf{v}_k,$$

qui vaut λ_k si $j = k$ et 0 sinon, puisque la base est orthonormée.

(3) Traitons d'abord le cas $q = 1$. Un vecteur unitaire $\mathbf{u}$ s'écrit $\mathbf{u} = c_1\mathbf{v}_1 + \cdots + c_p\mathbf{v}_p$, avec $c_j = \mathbf{u} \cdot \mathbf{v}_j$ et $c_1^2 + \cdots + c_p^2 = \|\mathbf{u}\|^2 = 1$. Alors $C\mathbf{u} = \lambda_1 c_1 \mathbf{v}_1 + \cdots + \lambda_p c_p \mathbf{v}_p$ et

$$\mathbf{u}^{\mathsf{T}} C\, \mathbf{u} = \lambda_1 c_1^2 + \lambda_2 c_2^2 + \cdots + \lambda_p c_p^2 \leq \lambda_1\,(c_1^2 + \cdots + c_p^2) = \lambda_1,$$

avec égalité pour $\mathbf{u} = \mathbf{v}_1$. La variance dans la direction $\mathbf{u}$ est donc une moyenne des valeurs propres, pondérée par les « poids » c_j^2, dont la somme vaut 1 ; elle est maximale lorsque tout le poids est placé sur la plus grande valeur propre.

Pour q quelconque, soit $(\mathbf{a}_1, \dots, \mathbf{a}_q)$ une base orthonormée d'un sous-espace V. La variance du nuage projeté sur V est $\sum_{l=1}^{q} \mathbf{a}_l^\intercal C\, \mathbf{a}_l$, et le calcul précédent donne

$$\sum_{l=1}^{q} \mathbf{a}_l^\intercal C\, \mathbf{a}_l = \sum_{j=1}^{p} \lambda_j\, w_j, \qquad \text{où} \qquad w_j = \sum_{l=1}^{q} (\mathbf{a}_l \cdot \mathbf{v}_j)^2.$$

Le poids w_j est le carré de la norme de la projection du vecteur unitaire $\mathbf{v}_j$ sur V : il est compris entre 0 et 1. De plus, $\sum_j w_j = \sum_l \sum_j (\mathbf{a}_l \cdot \mathbf{v}_j)^2 = \sum_l \|\mathbf{a}_l\|^2 = q$. Une somme $\sum_j \lambda_j w_j$ dont les poids sont compris entre 0 et 1 et totalisent q est au plus égale à $\lambda_1 + \cdots + \lambda_q$, la somme des q plus grandes valeurs propres : on obtient la plus grande somme en donnant le poids maximal, 1, aux q plus grandes valeurs propres. Cette borne est atteinte lorsque $V = \text{span}\{\mathbf{v}_1, \dots, \mathbf{v}_q\}$.

(4) Par le théorème de Pythagore, la moyenne des carrés des distances au sous-espace est la variance totale, $\lambda_1 + \cdots + \lambda_p$, moins la variance du nuage projeté. On conclut avec (3). □

Le point (2) est important en pratique. Les variables de départ sont souvent fortement corrélées entre elles, et donc redondantes : la longueur du bec et la longueur de la nageoire disent en partie la même chose. Les composantes principales, elles, sont non corrélées : chacune apporte une information que les précédentes ne contenaient pas.

Algorithme 5.1 **Analyse en composantes principales**

Partant d'un tableau de données de n individus et de p variables :

1. centrer chaque variable, en lui soustrayant sa moyenne ; si les variables ne sont pas mesurées dans la même unité, ou si leurs dispersions sont très différentes, les réduire aussi, en les divisant par leur écart type ;
2. calculer la matrice de covariance $C = \frac{1}{n} X^\intercal X$ du tableau obtenu ;
3. calculer les valeurs propres et des vecteurs propres unitaires de C, et les ranger par ordre décroissant des valeurs propres ;
4. choisir le nombre q d'axes à conserver ;
5. calculer les coordonnées des individus sur ces axes, $y_{ik} = \mathbf{x}_i \cdot \mathbf{v}_k$, et représenter le nuage projeté, par exemple dans le plan des deux premiers axes ;
6. interpréter les axes à l'aide des composantes des vecteurs propres.

5.4.3 Combien d'axes conserver ?

Il n'existe pas de règle universelle, mais trois critères sont d'usage courant. Le plus simple consiste à conserver assez d'axes pour expliquer une proportion fixée de la variance totale, par exemple 80 % ou 90 %. On peut aussi tracer le **diagramme des éboulis**, qui représente les valeurs propres par ordre décroissant (figure 5.9), et chercher un « coude » au-delà duquel les valeurs propres diminuent lentement : les axes suivants ne font qu'ajouter de petites contributions de même importance, qui relèvent souvent du bruit. Enfin, pour des données réduites, le **critère de Kaiser** conserve les axes dont la valeur propre dépasse 1. Puisque la variance totale vaut alors p et que chaque variable de départ y contribue pour 1, un axe dont la valeur propre est inférieure à 1 résume moins d'information qu'une seule des variables de départ.

Pour visualiser les données, la question est souvent tranchée d'avance : on veut un dessin,

donc $q = 2$ (ou $q = 3$, avec une figure que l'on peut faire tourner). Il faut alors garder en tête la proportion de variance conservée par le plan, qui indique jusqu'à quel point le dessin est fidèle aux données.

Remarque 5.5 **Le signe des axes est arbitraire**

Un vecteur propre n'est défini qu'à un facteur non nul près ; même unitaire, il peut être remplacé par son opposé (exercice 4.1). Changer $\mathbf{v}_k$ en $-\mathbf{v}_k$ change le signe de toutes les coordonnées y_{ik}, ce qui retourne la figure selon le k^{e} axe, sans rien changer à son interprétation. Deux logiciels, ou deux versions d'un même logiciel, peuvent donc produire des figures symétriques l'une de l'autre. Dans ce chapitre, nous avons choisi les signes qui rendent les figures plus faciles à lire, par exemple pour que les manchots papous soient à droite.

5.4.4 Le calcul avec Python

L'algorithme 5.1 se programme en quelques lignes avec `numpy`. La fonction `np.linalg.eigh`, réservée aux matrices symétriques (section D.11), renvoie les valeurs propres par ordre *croissant* ; on les remet dans l'ordre décroissant à l'aide de `np.argsort`. Pour lire le fichier de données, nous utilisons la bibliothèque `pandas`, l'outil de référence en Python pour manipuler des tableaux de données ; elle est installée d'office dans les environnements comme Google Colab. La méthode `dropna` élimine les deux manchots pour lesquels les mesures manquent.

Programme 5.1 : Une fonction pour l'ACP, appliquée aux trois mesures des manchots.

```
import numpy as np
import pandas as pd
np.set_printoptions(precision=3, suppress=True)

url = ("https://raw.githubusercontent.com/allisonhorst/"
       "palmerpenguins/main/inst/extdata/penguins.csv")
manchots = pd.read_csv(url).dropna(subset=["bill_length_mm"])
X = manchots[["bill_length_mm", "bill_depth_mm",
              "flipper_length_mm"]].to_numpy()

def acp(X, reduire=True):
    Z = X - X.mean(axis=0)              # centrer
    if reduire:
        Z = Z / X.std(axis=0)           # réduire : cotes z
    C = Z.T @ Z / len(Z)                # matrice de covariance
    valeurs, vecteurs = np.linalg.eigh(C)
    ordre = np.argsort(valeurs)[::-1]   # ordre décroissant
    valeurs, vecteurs = valeurs[ordre], vecteurs[:, ordre]
    return valeurs, vecteurs, Z @ vecteurs

valeurs, vecteurs, Y = acp(X)
print("taille :", X.shape)
print("valeurs propres :", valeurs)
print("proportions :", valeurs / valeurs.sum())
print(vecteurs)
print(Y[:3, :2])     # trois manchots dans le plan principal
```

Programme 5.2 : Sortie.

```
taille : (342, 3)
valeurs propres : [2.003 0.767 0.23 ]
proportions : [0.668 0.256 0.077]
[[-0.551  0.657 -0.514]
 [ 0.512  0.753  0.413]
 [-0.659  0.036  0.752]]
[[ 1.824 -0.041]
 [ 1.211 -0.476]
 [ 0.864 -0.128]]
```

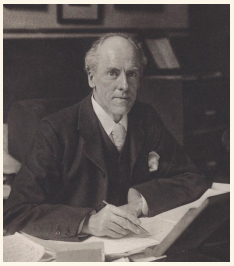
On retrouve les valeurs propres et les vecteurs propres annoncés à la section 5.3, au signe près du premier vecteur (remarque 5.5) : avec ce signe, les manchots papous se retrouvent à gauche. La dernière ligne du programme affiche les coordonnées des trois premiers manchots, tous de l'espèce Adélie, dans le plan principal. Il ne reste qu'à les dessiner avec `matplotlib`, en coloriant chaque point selon l'espèce (exercice 5.8).

Remarque 5.6 **L'ACP en pratique**

Les bibliothèques spécialisées, comme `scikit-learn`, permettent d'effectuer l'ACP sans avoir à programmer les calculs. Avec `PCA().fit(Z)`, l'attribut `explained_variance_ratio_` donne les proportions de variance expliquée, et la méthode `transform` calcule les composantes principales. Attention : l'attribut `explained_variance_` divise par $n - 1$ (remarque 5.2).

5.4.5 Un peu d'histoire

Note historique 5.1 **Karl Pearson et les « lignes et plans d'ajustement »**



Karl Pearson (1857–1936), en 1910. Domaine public, National Portrait Gallery (Londres), *Wikimedia Commons*.

Karl Pearson (1857–1936) est né à Londres. Après des études de mathématiques à l'Université de Cambridge, il étudia la physique, la philosophie et le droit en Allemagne, avant de devenir, en 1884, professeur de mathématiques appliquées et de mécanique à l'University College de Londres. Sous l'influence de Francis Galton et du zoologiste Walter Weldon, il se tourna vers l'étude statistique de l'hérédité et de l'évolution, qu'il appelait la *biométrie* [ORg].

Pearson est l'un des fondateurs de la statistique moderne. On lui doit l'expression « écart type », le coefficient de corrélation qui porte son nom, le test du khi-deux et la revue *Biometrika*, fondée en 1901. La même année, il publia un court article intitulé « Sur les lignes et les plans qui s'ajustent au plus près à des systèmes de points de l'espace » [Pea01]. Il y remarque que les deux droites de régression dépendent de la variable que l'on choisit d'expliquer, et il propose plutôt de chercher la droite, ou le plan, qui minimise la somme des carrés des distances *perpendiculaires* : c'est exactement le problème de la section 5.2. Il montre que cette droite passe par le centre de gravité du nuage et qu'elle est portée par le grand axe de l'« ellipsoïde de corrélation », c'est-à-dire par un vecteur propre de la matrice de covariance.

> Pearson fut aussi, comme Galton, un ardent défenseur de l'eugénisme, qui prétendait améliorer l'espèce humaine en contrôlant la reproduction ; en 1911, il devint le premier titulaire de la chaire d'eugénisme créée par le testament de Galton. Ces idées, qui ont servi à justifier des politiques de stérilisation forcée au xxe siècle, sont aujourd'hui unanimement rejetées ; en 2020, l'University College de Londres a retiré les noms de Galton et de Pearson de plusieurs de ses bâtiments. L'histoire de la statistique rappelle ainsi que des outils mathématiques solides peuvent être mis au service de projets condamnables.

La méthode fut redécouverte, indépendamment, par le statisticien et économiste américain Harold Hotelling (1895–1973), qui lui donna son nom dans un article de 1933 intitulé « Analyse d'un ensemble de variables statistiques en composantes principales » [Hot33]. Le contexte est celui de la psychologie de l'éducation : l'article paraît dans le *Journal of Educational Psychology*, et Hotelling y illustre la méthode sur des résultats d'élèves à des tests de lecture et d'arithmétique. Les psychologues cherchaient alors à expliquer les corrélations entre les résultats scolaires par un petit nombre de « facteurs » cachés. Charles Spearman avait observé en 1904 [Spe04] que les notes d'écoliers en lettres classiques, en français, en anglais et en mathématiques sont toutes corrélées positivement, et il en avait déduit l'existence d'un facteur général d'intelligence, qu'il notait g ; nos six élèves, dont l'axe principal mesure le « niveau général », en sont un écho lointain. Hotelling propose aussi une méthode itérative de calcul, voisine de la méthode des puissances du chapitre 4.

Pendant longtemps, l'ACP resta limitée par le calcul : trouver à la main les valeurs propres d'une matrice 10×10 est un travail de plusieurs jours. L'arrivée des ordinateurs rend ces calculs beaucoup plus accessibles. En France, l'école d'« analyse des données » de Jean-Paul Benzécri popularise, dans les années 1960 et 1970, l'ACP et ses représentations graphiques, qui restent très utilisées dans le monde francophone. La même idée apparaît aussi sous d'autres noms, comme la *transformation de Karhunen–Loève* en traitement du signal. Aujourd'hui, l'ACP est l'une des méthodes les plus employées de toute la statistique [Jol02 ; JC16].

5.5 Au-delà de trois dimensions

5.5.1 Le décathlon

L'intérêt de l'ACP apparaît vraiment lorsque les données ont plus de trois variables, et qu'aucun dessin direct n'est possible. Le décathlon en fournit un bel exemple. Cette épreuve d'athlétisme réunit dix disciplines, disputées sur deux jours : quatre courses (le 100 m, le 400 m, le 110 m haies et le 1500 m), trois sauts (en longueur, en hauteur et à la perche) et trois lancers (le poids, le disque et le javelot). Le tableau 5.3 donne un extrait de 41 résultats obtenus par des décathloniens lors de deux compétitions de 2004, le Décastar de Talence et les Jeux olympiques d'Athènes [HLP17]. Chaque résultat est un point de $\mathbb{R}^{10}$; quelques athlètes, comme Sebrle, Clay et Karpov, ont participé aux deux compétitions et apparaissent donc deux fois.

Les variables étant mesurées dans des unités différentes (secondes et mètres) et avec des dispersions très différentes, nous les réduisons ; la variance totale vaut donc 10. La fonction `acp` du programme 5.1 s'applique sans changement, une fois le fichier `decathlon.csv` lu. La colonne `Points` contient le total officiel des points de chaque athlète, que nous n'utilisons pas dans le calcul.

TABLE 5.3 : Les résultats de quelques décathloniens en 2004 : temps en secondes pour les courses, distances et hauteurs en mètres pour les sauts et les lancers.

Athlète	100	Long.	Poids	Haut.	400	Haies	Disque	Perche	Jav.	1500
Sebrle	10,85	7,84	16,36	2,12	48,36	14,05	48,72	5,00	70,52	280,01
Karpov	10,50	7,81	15,93	2,09	46,81	13,97	51,65	4,60	55,54	278,11
Drews	10,87	7,38	13,07	1,88	48,51	14,01	40,11	5,00	51,53	274,21
Casarsa	11,36	6,68	14,92	1,94	53,20	15,39	48,66	4,40	58,62	296,12
moyenne	11,00	7,26	14,48	1,98	49,62	14,61	44,33	4,76	58,32	279,02

Programme 5.3 : L'ACP du décathlon (la fonction `acp` est celle du programme 5.1).

```
deca = pd.read_csv("decathlon.csv", index_col=0)
X = deca.iloc[:, :10].to_numpy()        # les dix épreuves
valeurs, vecteurs, Y = acp(X)
print("valeurs propres :", valeurs.round(2))
print("% cumulés :", (100 * np.cumsum(valeurs) / valeurs.sum()).round(1)
    )
r = np.corrcoef(Y[:, 0], deca["Points"])[0, 1]
print("corrélation y1 / points :", round(r, 3))
```

Programme 5.4 : Sortie.

```
valeurs propres : [3.27 1.74 1.4  1.06 0.68 0.6  0.45 0.4  0.21 0.18]
% cumulés : [ 32.7  50.1  64.1  74.7  81.6  87.5  92.1  96.   98.2 100.
    ]
corrélation y1 / points : 0.956
```

Le diagramme des éboulis (figure 5.9) est beaucoup moins tranché que pour les manchots. Le premier axe explique 32,7 % de la variance, le deuxième 17,4 % : le plan principal ne conserve qu'environ la moitié de l'information. Le critère de Kaiser retiendrait quatre axes, et il en faudrait cinq pour dépasser 80 %. Ce n'est pas un échec de la méthode, mais une information sur les données : les performances dans les dix épreuves sont assez indépendantes les unes des autres, et l'on ne peut pas résumer un décathlonien par deux nombres sans perte importante. Le plan principal en donne néanmoins la meilleure image plane possible.

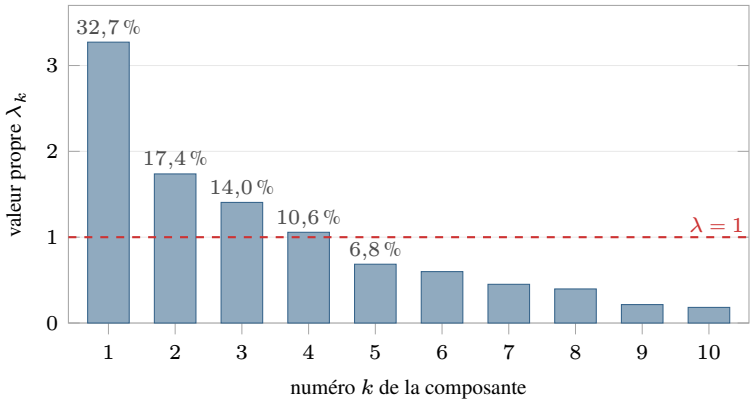


FIGURE 5.9 : Le diagramme des éboulis de l'ACP du décathlon, avec la proportion de variance expliquée par les cinq premiers axes. La droite pointillée indique le seuil du critère de Kaiser.

5.5.2 Lire le plan principal

La figure 5.10 montre les athlètes dans le plan des deux premiers axes. Pour interpréter ces axes, on regarde les composantes des vecteurs propres $\mathbf{v}_1$ et $\mathbf{v}_2$ (tableau 5.4). Sur le premier axe, les composantes des épreuves de vitesse (100 m, 110 m haies, 400 m) sont négatives, et celle du saut en longueur est positive. Comme un bon temps est un temps *court*, cela signifie que les athlètes situés à droite courent vite et sautent loin. Le premier axe mesure donc la performance générale, en particulier dans les épreuves de vitesse et de saut. Le second axe est porté par les lancers (poids et disque), mais aussi par le 400 m et le 1500 m : les athlètes du haut de la figure sont puissants, mais relativement lents sur les courses longues. C'est le cas de Casarsa, tout en haut, qui lance le disque à 48,66 m mais court le 400 m en 53,20 s, l'un des pires temps du groupe. Le saut à la perche, dont les deux composantes sont faibles, est mal représenté par ce plan : c'est le troisième axe qui le décrit.

TABLE 5.4 : Les composantes des deux premiers vecteurs propres de la matrice de corrélation du décathlon. Le signe de chaque vecteur est choisi pour que le saut en longueur soit du côté positif du premier axe, et le lancer du disque du côté positif du second.

Épreuve	$\mathbf{v}_1$	$\mathbf{v}_2$
100 m	−0,428	0,142
saut en longueur	0,410	−0,262
lancer du poids	0,344	0,454
saut en hauteur	0,316	0,266
400 m	−0,376	0,432
110 m haies	−0,413	0,174
lancer du disque	0,305	0,460
saut à la perche	0,028	−0,137
lancer du javelot	0,153	0,241
1500 m	−0,032	0,360

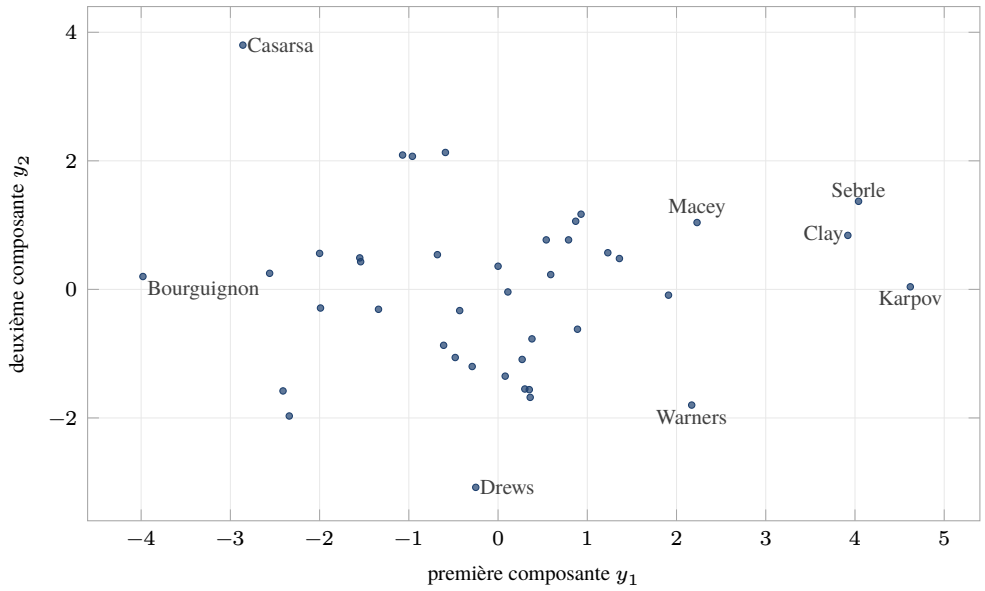


FIGURE 5.10 : Les 41 résultats du décathlon projetés sur le plan principal. Parmi les athlètes nommés, Bourguignon a concouru au Décastar, les autres aux Jeux olympiques.

Le résultat le plus frappant est celui de la dernière ligne du programme 5.3. La première composante principale a été calculée sans aucune connaissance du barème officiel

du décathlon, qui convertit chaque performance en points selon des tables complexes. Pourtant, son coefficient de corrélation avec le total des points est de 0,956 : les trois médaillés olympiques de 2004, Sebrle, Clay et Karpov, sont les trois points les plus à droite de la figure. L'ACP a, en quelque sorte, retrouvé seule ce que signifie « être un bon décathlonien ».

5.5.3 D'autres applications

Les mêmes calculs s'appliquent à des données de dimension bien plus grande, et l'ACP est souvent la première chose que l'on fait devant un nouveau jeu de données. En génétique, John Novembre et ses collègues ont appliqué l'ACP aux génomes de près de 1400 Européens, décrits par environ 200 000 marqueurs génétiques. Une fois chaque personne projetée sur le plan des deux premiers axes, et la figure convenablement tournée, le nuage reproduit de façon saisissante la carte géographique de l'Europe [Nov+08]. En reconnaissance d'images, chaque photographie de 100×100 pixels est un point de $\mathbb{R}^{10\,000}$; les vecteurs propres de la matrice de covariance d'un grand nombre de visages, que l'on peut afficher comme des images, ont été appelés *visages propres* (en anglais *eigenfaces*), et un visage se décrit assez bien par ses coordonnées sur quelques dizaines de ces visages propres [TP91]. L'ACP sert aussi à compresser des données, à éliminer du bruit, ou à réduire le nombre de variables avant d'appliquer une autre méthode d'apprentissage automatique.

Il faut toutefois garder en tête les limites de la méthode. L'ACP est *linéaire* : elle cherche des droites et des plans, et elle passe à côté d'une structure courbe, comme un nuage disposé le long d'un arc de cercle. Elle est sensible aux unités et aux valeurs aberrantes, qui peuvent accaparer un axe à elles seules (exercices 5.5 et 5.10). Enfin, une grande variance n'est pas toujours synonyme d'information intéressante : la direction qui sépare deux groupes peut très bien être une direction de faible variance.

Pour la classe 5.4 **Des données réelles en classe**

> Le programme de mathématique du secondaire insiste sur l'interprétation de données et sur le développement d'un regard critique sur l'information statistique [Min16]. Les données des manchots se prêtent bien à une activité en ce sens, sans aucune algèbre linéaire. On distribue aux élèves les mesures d'une cinquantaine de manchots, sans indiquer l'espèce, et on leur demande de construire des nuages de points pour différentes paires de variables : lequel montre le mieux qu'il y a « plusieurs sortes » de manchots ? La discussion fait apparaître l'idée centrale de ce chapitre : un même ensemble de données, regardé sous différents angles, peut révéler ou cacher sa structure. Avec des élèves plus avancés, un tableur permet de calculer la cote z de chaque mesure et de constater que les variables, une fois réduites, deviennent comparables. On peut aussi proposer aux élèves de construire un « score » combinant plusieurs épreuves sportives réalisées en éducation physique, puis de comparer leur classement à celui que produit une simple somme des cotes z.

Exercices

Exercice 5.1. On considère les quatre points $(1,1)$, $(3,3)$, $(3,7)$ et $(5,5)$ du plan.

a) Calculer le point moyen M et les coordonnées des points centrés.

b) Calculer la matrice de covariance C et le coefficient de corrélation r_{xy}.

c) Trouver les valeurs propres et des vecteurs propres unitaires de C.

d) Donner l'équation de l'axe principal du nuage, et la comparer avec celle de la droite de régression de y selon x. Faire un dessin.

e) Calculer les coordonnées t_i des projections des quatre points sur l'axe principal, et vérifier que leur variance est égale à la plus grande valeur propre.

f) Quelle proportion de la variance totale l'axe principal explique-t-il ? Calculer la somme des carrés des distances des quatre points à cet axe.

Exercice 5.2. Un nuage centré du plan a pour matrice de covariance $C = \begin{pmatrix} 13 & 3\sqrt{3} \\ 3\sqrt{3} & 7 \end{pmatrix}$.

a) Montrer que la variance des projections sur la droite dirigée par $\mathbf{u} = (\cos\theta, \sin\theta)$ est $f(\theta) = 10 + 3\cos(2\theta) + 3\sqrt{3}\,\sin(2\theta)$.

b) Écrire $f(\theta)$ sous la forme $10 + A\cos(2\theta - \varphi)$, et en déduire le maximum et le minimum de f, ainsi que les angles où ils sont atteints.

c) Vérifier le résultat en calculant les valeurs propres et les vecteurs propres de C.

Exercice 5.3. Dans une école, on a relevé les résultats d'un grand nombre d'élèves en mathématiques, en physique et en éducation physique, puis on les a réduits. La matrice de corrélation obtenue est

$$R = \begin{pmatrix} 1 & 0{,}8 & 0 \\ 0{,}8 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

a) Calculer les valeurs propres de R et des vecteurs propres unitaires. Vérifier que la somme des valeurs propres est égale à la variance totale.

b) Quelle proportion de la variance totale chaque axe explique-t-il ? Combien d'axes faut-il retenir pour en expliquer au moins $80\,\%$? Que dit le critère de Kaiser ?

c) Interpréter chacun des trois axes principaux.

d) Un élève a pour cotes z les valeurs $(1{,}5\,;\ 0{,}5\,;\ -1)$. Calculer ses coordonnées dans le plan des deux premiers axes, ainsi que le carré de la distance entre ce point et sa projection sur ce plan.

e) On ne conserve que les deux premiers axes. Quelles cotes z reconstruit-on pour cet élève ? Que représente le carré de la distance calculé en d) ? Le comparer à λ_3 et interpréter.

Exercice 5.4. On reprend les six élèves du tableau 5.1, mais on exprime maintenant la note de sciences sur 10 plutôt que sur 100 : on remplace y par $y' = y/10$.

a) Calculer la nouvelle matrice de covariance C' et le nouveau coefficient de corrélation.

b) Calculer les valeurs propres de C' et un vecteur propre associé à la plus grande. Quelle proportion de la variance totale le premier axe explique-t-il ?

c) Exprimer l'axe principal obtenu dans les unités de départ (notes sur 100) et le comparer à la droite $y = x + 5$ trouvée à la section 5.2. Conclure.

d) Montrer que si l'on réduit les variables, l'ACP ne dépend plus des unités choisies.

Exercice 5.5. On reprend les six élèves du tableau 5.1, et on leur ajoute deux élèves atypiques, G et H, dont les notes de mathématiques et de sciences sont $G = (70 + k,\ 75 - k)$ et $H = (70 - k,\ 75 + k)$, où $k > 0$.

a) Vérifier que le point moyen est toujours $(70, 75)$, et calculer la matrice de covariance C_k des huit élèves en fonction de k.

b) Montrer que, quel que soit k, les vecteurs propres de C_k sont encore $\mathbf{u} = \frac{1}{\sqrt{2}}(1, 1)$ et $\mathbf{w} = \frac{1}{\sqrt{2}}(1, -1)$, et calculer les valeurs propres associées.

c) On prend $k = 12$. Calculer le coefficient de corrélation, déterminer l'axe principal et la proportion de la variance qu'il explique. Comparer avec les résultats obtenus pour les six élèves.

d) À partir de quelle valeur de k l'axe principal bascule-t-il de $\mathbf{u}$ à $\mathbf{w}$? Que peut-on en conclure sur la sensibilité de l'ACP aux valeurs extrêmes ?

Exercice 5.6. Dire si chacun des énoncés suivants est vrai ou faux, et justifier brièvement.

a) La première composante principale est toujours l'une des variables de départ.

b) Si les variables de départ sont deux à deux non corrélées, l'ACP non réduite se contente de les ranger par ordre de variance décroissante.

c) Pour un nuage du plan, l'axe principal est la droite de régression de y selon x.

d) Pour des données réduites à p variables, la somme des valeurs propres vaut p.

e) Deux nuages qui ont la même matrice de covariance ont la même forme.

f) Si les deux premières composantes principales expliquent 95 % de la variance, deux individus proches dans le plan principal sont proches dans $\mathbb{R}^p$.



Exercice 5.7. On reprend les six élèves du tableau 5.1.

a) Avec `numpy`, calculer la matrice de covariance (avec `np.cov(X.T, bias=True)`, qui divise par n), puis ses valeurs propres et ses vecteurs propres.

b) Pour 181 valeurs de θ régulièrement réparties entre 0° et 180°, calculer la variance des projections des points centrés sur la droite dirigée par $(\cos\theta, \sin\theta)$, et tracer le graphique de la figure 5.4. Pour quel angle le maximum est-il atteint ?

c) Avec `np.polyfit(x, y, 1)`, calculer la droite de régression de y selon x et vérifier que sa pente vaut s_{xy}/s_x^2. Tracer le nuage, l'axe principal et les deux droites de régression.

d) Échanger les rôles de x et de y. Lesquelles des trois droites changent ?



Exercice 5.8. On reprend les données des manchots et la fonction `acp` du programme 5.1.

a) Reproduire la figure 5.7 avec `matplotlib`, en coloriant chaque point selon l'espèce (colonne `species`).

b) Reproduire la figure 5.8. Les manchots papous sont-ils à gauche ou à droite ? Pourquoi est-ce sans importance ?

c) Recommencer l'ACP *sans* réduire les variables (`acp(X, reduire=False)`). Quelle proportion de la variance le premier axe explique-t-il ? Quelle variable domine le premier vecteur propre ? Expliquer.

d) Vérifier numériquement que les trois composantes principales de l'ACP réduite sont non corrélées et que leurs variances sont les valeurs propres.



Exercice 5.9. On ajoute aux trois mesures des manchots une quatrième variable, la masse en grammes (colonne `body_mass_g`).

a) Effectuer l'ACP de ces quatre variables sans les réduire. Que constate-t-on ?

b) Effectuer l'ACP des variables réduites. Donner les proportions de variance expliquée. Combien d'axes retient le critère de Kaiser ? combien en faut-il pour dépasser 80 % ?

c) Interpréter le premier axe à l'aide de son vecteur propre.

d) Reconstruire chaque manchot à partir de ses deux premières composantes, $\hat{\mathbf{z}}_i = y_{i1}\mathbf{v}_1 + y_{i2}\mathbf{v}_2$, et vérifier que la moyenne des $\|\mathbf{z}_i - \hat{\mathbf{z}}_i\|^2$ est égale à $\lambda_3 + \lambda_4$ (théorème 5.4).

Exercice 5.10. On reprend les quatre variables de l'exercice 5.9 (tableau `X4`). Lors de la saisie des données, on a inscrit par erreur une masse de 42 000 g au lieu de 4200 g pour le manchot de la ligne 18 (un manchot Adélie de l'île Torgersen).

a) Introduire cette erreur dans une copie du tableau : `X4e = X4.copy()` puis `X4e[18, 3] = 42000`. Comparer l'écart type de la masse, ainsi que son coefficient de corrélation avec la longueur de la nageoire, avant et après l'erreur.

b) Effectuer l'ACP réduite des données erronées. Comparer les valeurs propres, les proportions de variance expliquée et les deux premiers vecteurs propres avec ceux de l'exercice 5.9.

c) Calculer les coordonnées de ce manchot dans le plan principal, et les comparer à celles des autres manchots.

d) Calculer les cotes z de toutes les mesures. Comment cette erreur aurait-elle pu être détectée avant l'ACP ? Pourquoi la réduction des variables ne protège-t-elle pas contre une telle erreur ?

Exercice 5.11. La bibliothèque `scikit-learn` contient un jeu de données classique décrivant 178 vins italiens, issus de trois cépages, par 13 variables chimiques (taux d'alcool, acidité, intensité de la couleur, etc.) [AF91]. On le charge avec

```
from sklearn.datasets import load_wine
vins = load_wine()
X, cepage = vins.data, vins.target   # cepage vaut 0, 1 ou 2
```

a) Pourquoi faut-il réduire les variables ? (Comparer les écarts types des variables.)

b) Effectuer l'ACP réduite avec la fonction `acp`, tracer le diagramme des éboulis et donner la proportion de variance conservée par le plan principal.

c) Représenter les vins dans le plan principal, en coloriant chacun selon son cépage. Que constate-t-on ?

d) Comparer avec la classe `PCA` de `scikit-learn`, appliquée aux données réduites par `StandardScaler`. Les proportions de variance sont-elles les mêmes ? et les valeurs propres ?

Exercice 5.12. L'analyse en composantes principales ne figure pas au programme du secondaire, mais elle mobilise plusieurs notions qui y sont étudiées, et elle intervient de plus en plus dans les médias, sous forme de graphiques qui « résument » des données complexes.

a) Relever trois notions du programme du secondaire [Min16] qui interviennent dans ce chapitre, et préciser le rôle de chacune.

b) Concevoir les grandes lignes d'une activité d'une période, destinée à des élèves du deuxième cycle, qui fasse vivre l'idée de « choisir le meilleur point de vue » sur un ensemble de données, sans algèbre linéaire.

c) Un article présente un graphique dans lequel des pays sont placés selon deux axes nommés « développement économique » et « qualité de vie », obtenus par une ACP de vingt indicateurs. Quelles questions un lecteur critique devrait-il se poser ? Comment l'enseignant peut-il amener les élèves à se les poser ?

6 Numération et stockage de l'information

« Dans un ordinateur, tout est fait de 0 et de 1 : le courant passe ou ne passe pas. » Cette explication, que tout le monde a entendue, n'est pas fausse, mais elle laisse de côté presque tout ce qui est intéressant. Comment une suite de 0 et de 1 peut-elle devenir la lettre « é », le rouge d'un logo ou la photo d'un visage ? Pourquoi écrit-on les couleurs d'une page Web sous la forme bizarre `#1C3F6E` ? Pourquoi un disque dur vendu pour « 200 Go » n'en affiche-t-il que 186 ? Et où, physiquement, ces 0 et ces 1 se trouvent-ils ?

La réponse à ces questions repose sur une idée que nous enseignons depuis l'école primaire : la *numération positionnelle*. Écrire 2026, c'est dire $2 \times 1000 + 0 \times 100 + 2 \times 10 + 6$; rien n'oblige à grouper par dix, et l'ordinateur groupe par deux. Ce chapitre part donc de la numération en base dix, que nous généraliserons à une base quelconque. Nous étudierons ensuite le *binaire*, la langue naturelle des circuits, puis deux bases qui servent d'abréviations au binaire : l'*octal* et l'*hexadécimal*, dont les codes de couleur du Web sont l'exemple le plus parlant. Nous verrons alors comment le texte est encodé, de l'ASCII jusqu'à l'UTF-8, l'encodage utilisé aujourd'hui par la quasi-totalité du Web.

Dans la seconde moitié du chapitre, nous mesurerons l'information. Nous définirons l'octet et ses multiples (et nous comprendrons enfin où sont passés les gigaoctets manquants), nous distinguerons la mémoire vive du disque dur, et nous verrons comment un bit est physiquement inscrit dans un circuit, sur un disque magnétique ou dans une clé USB. Un bref historique des capacités de stockage montrera une croissance exponentielle comme on en voit rarement. Nous terminerons par les images numériques, en distinguant les images *matricielles*, faites de pixels, des images *vectorielles*, faites d'objets géométriques.

6.1 Écrire un nombre dans une base quelconque

6.1.1 La numération en base dix

Dans l'écriture 2026, le chiffre 2 apparaît deux fois, mais il ne représente pas la même quantité : à gauche, il compte des milliers, à droite, des dizaines. C'est le principe d'une **numération positionnelle** : la valeur d'un chiffre dépend de sa position.

$$2026 = 2 \times 10^3 + 0 \times 10^2 + 2 \times 10^1 + 6 \times 10^0.$$

Ce système, venu de l'Inde et transmis à l'Europe par les mathématiciens de langue arabe, n'utilise que dix symboles, de 0 à 9, et le zéro y joue un rôle essentiel : il marque une position vide. Le nombre 10 n'a rien de spécial, si ce n'est que nous avons dix doigts. D'autres civilisations ont fait d'autres choix : les Babyloniens comptaient en base soixante, ce qui nous vaut encore nos 60 minutes et nos 60 secondes ; les Mayas comptaient en base vingt, comme en témoigne peut-être le « quatre-vingts » du français.

6.1.2 Une base quelconque

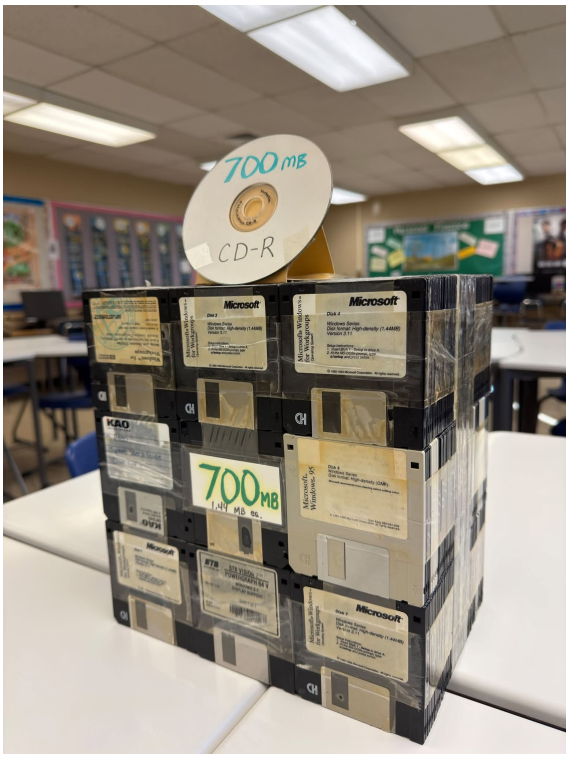


FIGURE 6.1 : Environ 500 disquettes de « 1,44 Mo » contiennent autant d'information qu'un seul CD-R de « 700 Mo ». Nous verrons à la section 6.5 que ces deux « Mo » ne désignent pas la même unité, et que le compte tombe pile : $737\,280\,000 \div 1\,474\,560 = 500$. Photo : u/itsthewolfe, « Visual representation of data, 486 floppies to 1 CD », *Reddit* (r/Damnthatsinteresting).

Définition 6.1 **Écriture en base b**

Soit $b \geq 2$ un entier, appelé la **base**. On dit que $c_{k-1}c_{k-2}\cdots c_1c_0$ est l'**écriture en base b** de l'entier naturel n si chaque **chiffre** c_i est un entier tel que $0 \leq c_i \leq b-1$, si $c_{k-1} \neq 0$ et si

$$n = c_{k-1}\,b^{k-1} + c_{k-2}\,b^{k-2} + \cdots + c_1\,b + c_0 = \sum_{i=0}^{k-1} c_i\,b^i.$$

On note alors $n = (c_{k-1}\cdots c_1c_0)_b$. Le chiffre c_0 est le chiffre **de poids faible**, et c_{k-1}, le chiffre **de poids fort**.

Par exemple, $(1011)_2 = 1\cdot 8 + 0\cdot 4 + 1\cdot 2 + 1 = 11$ et $(213)_5 = 2\cdot 25 + 1\cdot 5 + 3 = 58$. Lorsqu'aucune base n'est indiquée, il s'agit de la base dix. Pour passer d'une écriture en base b à la base dix, il suffit donc d'évaluer une somme. On peut réduire le nombre de multiplications en regroupant les termes comme suit (c'est le *schéma de Horner*) :

$$(213)_5 = (2\cdot 5 + 1)\cdot 5 + 3 = 11\cdot 5 + 3 = 58.$$

On lit les chiffres de gauche à droite et, à chaque nouveau chiffre, on multiplie le résultat partiel par la base avant d'ajouter ce chiffre.

Le passage inverse, de la base dix à la base b, repose sur la division euclidienne (théorème E.2). Le dernier chiffre c_0 est le reste de la division de n par b, puisque

$$n = \left(c_{k-1}\,b^{k-2} + \cdots + c_1\right)b + c_0 \quad \text{avec} \quad 0 \leq c_0 < b,$$

et le quotient $\left(c_{k-1}\cdots c_1\right)_b$ s'écrit avec les chiffres restants. On recommence donc avec le quotient, jusqu'à obtenir un quotient nul.

Algorithme 6.1 **Écriture en base b par divisions successives**

Partant d'un entier $n \geq 1$:

1. effectuer la division euclidienne de n par b : $n = qb + r$, avec $0 \leq r < b$;
2. le reste r est le chiffre suivant, en allant de droite à gauche ;
3. si $q \neq 0$, remplacer n par q et retourner à l'étape 1 ; sinon, l'écriture est terminée.

Exemple 6.1 **L'année 2026 en base deux**

Les divisions successives par 2 donnent

$$\begin{aligned}
2026 &= 2\times 1013 + 0, & 1013 &= 2\times 506 + 1, & 506 &= 2\times 253 + 0,\\
253 &= 2\times 126 + 1, & 126 &= 2\times 63 + 0, & 63 &= 2\times 31 + 1,\\
31 &= 2\times 15 + 1, & 15 &= 2\times 7 + 1, & 7 &= 2\times 3 + 1,\\
3 &= 2\times 1 + 1, & 1 &= 2\times 0 + 1.
\end{aligned}$$

En lisant les restes du dernier au premier, on obtient $2026 = (11111101010)_2$. On vérifie : $1024 + 512 + 256 + 128 + 64 + 32 + 8 + 2 = 2026$.

L'algorithme 6.1 montre que tout entier naturel possède une écriture en base b. Il en possède une seule.

Théorème 6.1 **Existence et unicité de l'écriture en base b**

Soit $b \geq 2$. Tout entier $n \geq 1$ possède une et une seule écriture en base b.

Nous admettrons ce résultat. L'algorithme 6.1 fournit une écriture, et l'on peut montrer qu'il n'y en a pas d'autre, parce que le reste d'une division euclidienne est unique.

Combien faut-il de chiffres pour écrire n en base b ? Un nombre s'écrit avec k chiffres exactement lorsque $b^{k-1} \leq n < b^k$, c'est-à-dire lorsque $k - 1 \leq \log_b n < k$. Ainsi, le nombre de chiffres de n en base b est $\lfloor \log_b n \rfloor + 1$. Plus la base est petite, plus l'écriture est longue : comme $\log_2 n = \log_2 10 \cdot \log_{10} n \approx 3{,}32 \log_{10} n$, il faut environ 3,3 fois plus de chiffres en base deux qu'en base dix. C'est le prix à payer pour n'utiliser que deux symboles.

Pour la classe 6.1 **Les bases au secondaire**

La numération positionnelle est enseignée dès le primaire, à l'aide du matériel de base dix, des abaques et des tableaux de numération. Au secondaire, les élèves rencontrent les puissances, la notation exponentielle et, en quatrième et cinquième secondaire (séquences TS et SN), les fonctions exponentielles et logarithmiques [Min16]. Le changement de base est un excellent terrain pour réinvestir ces notions : faire écrire des nombres en base cinq avec des jetons et des « paquets de cinq » oblige les élèves à expliciter ce que la base dix leur a rendu invisible, et la question « combien de chiffres ? » mène directement au logarithme.

6.2 Le binaire, langue des circuits

6.2.1 Pourquoi deux symboles ?

Un ordinateur ne manipule pas des 0 et des 1, mais des grandeurs physiques : une tension électrique dans un circuit, l'orientation de l'aimantation d'une minuscule région d'un disque, la quantité de charge piégée dans une cellule de mémoire flash. Ces grandeurs varient de façon continue et elles sont soumises au bruit : une tension n'est jamais exactement de 5 volts, elle fluctue avec la température, les interférences, le vieillissement des composants.

Pour représenter de l'information de façon fiable, on découpe la plage des valeurs possibles en quelques zones bien séparées. Dans un circuit alimenté en 5 volts, par exemple, une tension inférieure à 0,8 volt est lue comme un 0, une tension supérieure à 2 volts est lue comme un 1, et la zone intermédiaire est interdite : un circuit bien conçu n'y laisse jamais séjourner un signal. Avec deux zones seulement, la marge de sécurité est très grande : il faudrait une perturbation énorme pour faire passer un signal d'une zone à l'autre. Avec dix zones, pour représenter directement les chiffres de 0 à 9, chacune n'aurait qu'un demi-volt de largeur, et le même bruit ferait déborder le signal sur les zones voisines : un 7 serait lu comme un 6 ou un 8 (figure 6.2).

Le choix de la base deux n'est donc pas d'abord un choix mathématique, mais un compromis d'ingénierie : on sacrifie la compacité de l'écriture au profit de la *robustesse*. L'expression « le courant passe ou ne passe pas » en est une simplification ; il serait plus juste de dire « la tension est basse ou elle est haute ».

Une deuxième raison tient à la logique. En 1937, dans son mémoire de maîtrise au

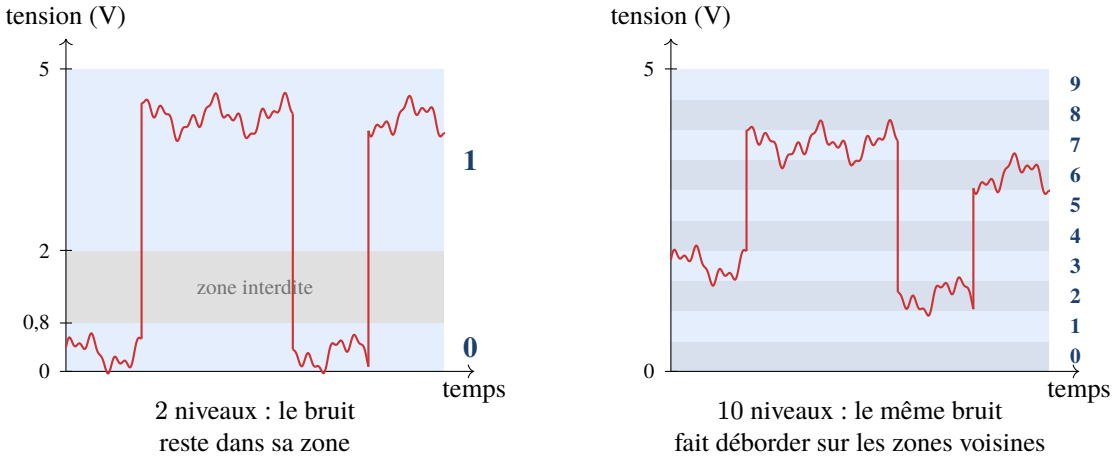


FIGURE 6.2 : Le même bruit électrique, de quelques dixièmes de volt, appliqué à un signal à deux niveaux (à gauche) et à un signal à dix niveaux (à droite). Avec deux niveaux, chaque valeur reste loin des seuils ; avec dix niveaux, les zones sont si étroites que le signal déborde constamment sur les zones voisines.

MIT, Claude Shannon montre que les circuits faits d'interrupteurs et de relais obéissent exactement aux règles de l'algèbre de Boole, où les variables ne prennent que deux valeurs, vrai ou faux [Sha38]. Les opérations logiques « et », « ou » et « non » se réalisent facilement avec des transistors, et l'on peut construire avec elles un additionneur, puis un processeur complet. Le binaire est ainsi à la fois une façon d'écrire les nombres et une façon de raisonner.

Définition 6.2 **Bit**

Un **bit** (contraction de l'anglais *binary digit*) est un chiffre de la base deux : il vaut 0 ou 1. Une **chaîne de n bits** est une suite de n bits.

Par le principe multiplicatif, il existe 2^n chaînes de n bits : chacun des n bits peut prendre deux valeurs, indépendamment des autres. Une chaîne de n bits peut donc représenter 2^n objets différents, par exemple les entiers de 0 à $2^n - 1$. Réciproquement, pour distinguer N objets, il faut au moins $\lceil \log_2 N \rceil$ bits : pour les 26 lettres minuscules, 5 bits suffisent ($2^5 = 32$), mais pas 4 ($2^4 = 16$).

Note historique 6.1 **Leibniz et l'arithmétique binaire**

Le binaire est bien plus ancien que l'ordinateur. Gottfried Wilhelm Leibniz (1646–1716) en publie une description complète en 1703 dans un mémoire présenté à l'Académie royale des sciences de Paris, *Explication de l'arithmétique binaire, qui se sert des seuls caractères 0 et 1* [Lei03]. Il y donne les tables d'addition et de multiplication, remarquablement simples, et voit dans ce système une image de la création du monde à partir du néant (le 0) et de l'unité (le 1). Il croit aussi le reconnaître dans les hexagrammes du *Yi King*, un ancien texte chinois que lui avait fait connaître le jésuite Joachim Bouvet. Leibniz avait par ailleurs conçu une machine à calculer, mais il n'imaginait pas qu'une machine pût calculer en binaire. Il faudra attendre George Boole (1847), Claude Shannon (1937) et les premiers ordinateurs électroniques des années 1940 pour que le binaire devienne la langue

> des machines. Le mot *bit* lui-même apparaît en 1948 dans l'article fondateur de la théorie de l'information, où Shannon en attribue l'invention à son collègue John Tukey [Sha48].

6.2.2 Calculer en binaire

Les algorithmes appris à l'école pour la base dix fonctionnent dans toutes les bases, avec des tables beaucoup plus courtes en base deux : $0+0=0$, $0+1=1+0=1$ et $1+1=(10)_2$, c'est-à-dire 0 avec une retenue de 1. Par exemple,

$$\begin{array}{r} 1\,0\,1\,1 \\ +\;\; 0\,1\,1\,1 \\ \hline 1\,0\,0\,1\,0 \end{array} \qquad \text{soit} \qquad 11+7=18.$$

La multiplication par 2 est encore plus simple : comme la multiplication par 10 en base dix, elle consiste à ajouter un 0 à droite, c'est-à-dire à *décaler* tous les bits d'un rang vers la gauche. Les processeurs utilisent abondamment ce décalage, beaucoup plus rapide qu'une multiplication générale. De même, le reste de la division par 2^k se lit directement : ce sont les k derniers bits.

6.2.3 Des entiers de taille fixe

Un processeur ne manipule pas des chaînes de longueur arbitraire, mais des **mots** de taille fixe : 8, 16, 32 ou 64 bits. Avec 8 bits, on représente les entiers de 0 à 255 ; avec 32 bits, de 0 à $2^{32}-1=4\,294\,967\,295$. Que se passe-t-il si un calcul dépasse cette limite ? Le bit de retenue qui « déborde » à gauche est tout simplement perdu. Sur 8 bits, $200+100$ donne $300-256=44$. Autrement dit, l'arithmétique sur n bits est l'arithmétique *modulo* 2^n (section E.6).

Ce débordement s'appelle un **dépassement de capacité**, et ce n'est pas qu'une curiosité. Pour représenter aussi les entiers négatifs, on réserve en général un des bits au signe ; un entier signé de 32 bits va donc de -2^{31} à $2^{31}-1=2\,147\,483\,647$. C'est ainsi que YouTube comptait les visionnements de chaque vidéo. En décembre 2014, le vidéoclip *Gangnam Style* du chanteur sud-coréen Psy s'approchait de cette limite, et YouTube a dû passer à des entiers de 64 bits, qui permettent de compter jusqu'à environ $9{,}2\times10^{18}$. De nombreux systèmes comptent le temps en secondes écoulées depuis le 1er janvier 1970 dans un entier signé de 32 bits : ce compteur débordera le 19 janvier 2038, ce qu'on appelle le « bogue de l'an 2038 ».

6.2.4 Et les nombres à virgule ?

Les nombres réels posent un problème plus profond. En base dix, la fraction $\frac{1}{3}=0{,}333\ldots$ n'a pas de développement fini : au secondaire, on apprend qu'une fraction irréductible $\frac{p}{q}$ a un développement décimal fini si et seulement si q n'a pas d'autres facteurs premiers que 2 et 5, les facteurs premiers de 10. Le même raisonnement vaut en base deux : $\frac{p}{q}$ a un développement binaire fini si et seulement si q est une puissance de 2. Or $\frac{1}{10}$ n'est pas dans ce cas, et

$$0{,}1=(0{,}0\,0011\,0011\,0011\ldots)_2,$$

avec une période infinie. Un ordinateur, qui stocke les nombres sur un nombre fini de bits (section 6.2.5), doit donc *arrondir* 0,1. D'où le résultat surprenant, en Python :

```
>>> 0.1 + 0.2
0.30000000000000004
```

```
>>> 0.1 + 0.2 == 0.3
False
```

L'erreur est minuscule, mais elle peut s'accumuler [Gol91].

6.2.5 La norme IEEE 754

Jusqu'au début des années 1980, chaque fabricant d'ordinateurs représentait les nombres à virgule à sa façon : nombre de bits, règles d'arrondi, traitement des divisions par zéro, tout variait, et un même programme pouvait donner des résultats différents d'une machine à l'autre. Pour y mettre de l'ordre, l'IEEE (*Institute of Electrical and Electronics Engineers*), une association internationale d'ingénieurs qui publie de nombreuses normes techniques (celle du Wi-Fi, par exemple, est la norme IEEE 802.11), forme en 1977 un comité chargé d'une norme pour l'arithmétique des nombres à virgule. Le mathématicien William Kahan, de l'université de Californie à Berkeley, en est l'architecte principal. La norme IEEE 754 est adoptée en 1985, puis révisée en 2008 et en 2019.

Une norme n'est pas une loi. C'est un document technique rédigé par un comité d'experts (ici, des représentants des fabricants de processeurs et des chercheurs en analyse numérique), discuté et amendé pendant des années, puis adopté par un vote. Les fabricants la suivent volontairement, parce qu'un processeur qui la respecte peut exécuter les programmes écrits pour tous les autres. Kahan a reçu le prix Turing en 1989, en grande partie pour ce travail. Aujourd'hui, pratiquement tous les processeurs, des téléphones aux superordinateurs, appliquent la norme IEEE 754.

L'idée est celle de la notation scientifique, mais en base deux. Tout réel x non nul s'écrit

$$x = \pm\,(1,f)_2 \times 2^{E},$$

où la **mantisse** $(1,f)_2$ est comprise entre 1 et 2, et où l'exposant E est un entier. Comme le chiffre avant la virgule est toujours 1, il est inutile de le stocker. En *double précision*, le format le plus courant (c'est celui des nombres à virgule de Python), un nombre occupe 64 bits, répartis en trois champs :

- 1 bit de signe s, qui vaut 0 pour un nombre positif et 1 pour un nombre négatif ;
- 11 bits pour l'exposant, sous la forme $e = E + 1023$; ce décalage évite d'avoir à représenter des exposants négatifs, puisque e est toujours positif ;
- 52 bits pour la partie f de la mantisse qui suit la virgule.

Ainsi, $x = (-1)^s \times (1,f)_2 \times 2^{e-1023}$. La *simple précision* suit le même principe sur 32 bits : 1 bit de signe, 8 bits d'exposant (décalé de 127) et 23 bits de mantisse. Le partage des bits entre l'exposant et la mantisse est un choix du comité : plus de bits d'exposant donnent des nombres plus grands et plus petits, plus de bits de mantisse donnent plus de précision.

Exemple 6.2 **Le nombre $-6{,}25$ en double précision**

Comme $6{,}25 = 4 + 2 + \frac{1}{4} = (110{,}01)_2$, on déplace la virgule de deux rangs vers la gauche : $6{,}25 = (1{,}1001)_2 \times 2^2$. Le signe est $s = 1$, l'exposant stocké est $e = 2 + 1023 = 1025 = (10000000001)_2$, et la partie f est `1001`, complétée par des 0 jusqu'à 52 bits :

$$\underbrace{1}_{s}\ \underbrace{10000000001}_{e\ (11\text{ bits})}\ \underbrace{1001\,0000\,0000\cdots0000}_{f\ (52\text{ bits})}.$$

Ce nombre est représenté exactement, parce que son développement binaire est fini.

Pour 0,1, c'est une autre histoire. On a $0{,}1 \times 16 = 1{,}6$, d'où $0{,}1 = (1{,}1001\,1001\,1001\,...)_2 \times 2^{-4}$ et $e = 1019$; mais la partie f est infinie, et il faut la couper après 52 bits. La norme impose d'arrondir *au plus proche*. Le nombre réellement stocké est

$$0{,}100\,000\,000\,000\,000\,005\,551\,115\,123\,125\,782\,7\,\ldots\,,$$

une erreur d'environ $5{,}6 \times 10^{-18}$. Plus généralement, avec 52 bits après la virgule, l'erreur relative d'arrondi est au plus $2^{-53} \approx 1{,}1 \times 10^{-16}$: la double précision garantit environ 15 à 16 chiffres décimaux significatifs. Les exposants vont de -1022 à 1023, si bien que le plus grand nombre représentable est environ $1{,}8 \times 10^{308}$. Les deux valeurs extrêmes du champ d'exposant sont réservées : $e = 0$ sert à représenter 0 et des nombres extrêmement petits, et $e = 2047$ sert à représenter l'infini (résultat de $1/0$, par exemple) et la valeur spéciale *NaN* (*Not a Number*, résultat de $0/0$). Grâce à ces conventions, un calcul impossible ne fait pas planter la machine : il produit une valeur qui signale le problème.

Remarque 6.1 **Un dixième de seconde mal arrondi**

Le 25 février 1991, pendant la guerre du Golfe, une batterie de missiles Patriot à Dhahran, en Arabie saoudite, n'a pas intercepté un missile Scud, qui a tué 28 soldats. Le système comptait le temps en dixièmes de seconde, et convertissait ce compteur en secondes en le multipliant par 0,1, stocké en binaire sur 24 bits. L'erreur d'arrondi, d'environ 0,000 000 095 seconde à chaque dixième de seconde, s'accumulait : après 100 heures de fonctionnement sans redémarrage, l'horloge interne était décalée d'environ 0,34 seconde. Un Scud parcourt plus d'un demi-kilomètre dans ce laps de temps, et le radar a cherché le missile au mauvais endroit [Uni92]. On pourra refaire le calcul à l'exercice 6.9.

6.3 L'octal et l'hexadécimal, abréviations du binaire

6.3.1 Le problème de la lisibilité

Le binaire convient aux circuits, mais très mal aux humains. Qui pourrait retenir, ou même recopier sans erreur, l'adresse `11000000101010000000000010000001` ? On pourrait la convertir en base dix, mais la conversion est laborieuse et, surtout, elle masque les bits : en regardant 1 616 904 321, on ne voit pas quels bits valent 1. Les bases huit et seize offrent un compromis idéal, grâce à la propriété suivante.

Proposition 6.2 **Conversion par regroupement**

Soit $m \geq 1$. Pour écrire un entier en base 2^m, il suffit de grouper les bits de son écriture binaire par blocs de m, en partant de la droite (en complétant au besoin par des 0 à gauche), et de remplacer chaque bloc par le chiffre de la base 2^m qu'il représente.

Plutôt que de démontrer cette proposition en général, voyons pourquoi elle fonctionne sur un exemple, en base $4 = 2^2$, où l'on groupe les bits deux par deux. Reprenons $2026 = (11111101010)_2$ (exemple 6.1) :

$$2026 = 2^{10} + 2^9 + 2^8 + 2^7 + 2^6 + 2^5 + 2^3 + 2^1.$$

Groupons les puissances de 2 deux par deux, en partant de la droite, et mettons en évidence

une puissance de 4 dans chaque groupe, à l'aide de $2^{2j} = 4^j$ et $2^{2j+1} = 2 \times 4^j$:

$$\begin{aligned} 2026 &= 2^{10} + (2^9 + 2^8) + (2^7 + 2^6) + (2^5 + 0) + (2^3 + 0) + (2^1 + 0) \\ &= 1 \times 4^5 + (2+1) \times 4^4 + (2+1) \times 4^3 + (2+0) \times 4^2 + (2+0) \times 4 + (2+0) \\ &= 1 \times 4^5 + 3 \times 4^4 + 3 \times 4^3 + 2 \times 4^2 + 2 \times 4 + 2. \end{aligned}$$

Chaque coefficient est compris entre 0 et 3 : c'est un chiffre de la base quatre, et c'est exactement la valeur du bloc de deux bits correspondant ($(01)_2 = 1$, $(11)_2 = 3$, $(10)_2 = 2$). Par l'unicité de l'écriture en base quatre (théorème 6.1),

$$2026 = (01\ 11\ 11\ 10\ 10\ 10)_2 = (133222)_4.$$

Le même raisonnement s'applique à n'importe quel entier et à n'importe quelle base 2^m : un bloc de m bits consécutifs, à partir du rang mj, vaut un nombre compris entre 0 et $2^m - 1$, multiplié par $2^{mj} = (2^m)^j$. L'exercice 6.5 applique la proposition aux bases huit et seize.

Ainsi, chaque chiffre octal ($8 = 2^3$) correspond exactement à 3 bits, et chaque chiffre hexadécimal ($16 = 2^4$), à 4 bits. La conversion se fait de tête, bloc par bloc, sans aucune division. Rien de tel n'est possible entre la base deux et la base dix, parce que 10 n'est pas une puissance de 2.

6.3.2 L'octal

En base huit, les chiffres vont de 0 à 7. Par exemple,

$$2026 = (011\ 111\ 101\ 010)_2 = (3752)_8.$$

L'octal a été très utilisé dans les années 1960 et 1970, sur des ordinateurs dont la taille des mots était un multiple de 3 : 12 bits pour le PDP-8, 36 bits pour plusieurs grands ordinateurs d'IBM et de DEC. Il a depuis été largement supplanté par l'hexadécimal, mais il survit dans un usage où les blocs de 3 bits ont un sens naturel : les permissions des fichiers dans les systèmes Unix, comme Linux et macOS.

Chaque fichier possède trois permissions, la lecture (*r*, pour *read*), l'écriture (*w*) et l'exécution (*x*), attribuées séparément à trois catégories d'utilisateurs : le propriétaire, son groupe et tous les autres. Cela fait $3 \times 3 = 9$ bits, que l'on groupe naturellement par trois. La commande `chmod 755 fichier` signifie :

$$(755)_8 = (\underbrace{111}_{rwx}\ \underbrace{101}_{r\text{-}x}\ \underbrace{101}_{r\text{-}x})_2,$$

c'est-à-dire que le propriétaire peut tout faire, et que les autres peuvent lire et exécuter le fichier, mais pas le modifier. Chaque chiffre octal est la somme des permissions accordées : 4 pour la lecture, 2 pour l'écriture, 1 pour l'exécution.

6.3.3 L'hexadécimal

En base seize, il faut seize chiffres. On complète les dix chiffres usuels par les six premières lettres de l'alphabet : $A = 10$, $B = 11$, $C = 12$, $D = 13$, $E = 14$ et $F = 15$.

L'intérêt de l'hexadécimal tient en une phrase : *un octet, c'est-à-dire un groupe de* 8 *bits, s'écrit toujours avec exactement deux chiffres hexadécimaux*, de `00` à `FF`. Par exemple,

$$(1110\ 1001)_2 = (E9)_{16} = 14 \times 16 + 9 = 233.$$

Comme l'octet est l'unité de base de la mémoire des ordinateurs (section 6.5), l'hexadécimal est devenu la notation universelle des informaticiens pour afficher le contenu brut

TABLE 6.1 : Les seize chiffres hexadécimaux, leur valeur et leur écriture sur 4 bits.

Hexa	Décimal	Binaire	Hexa	Décimal	Binaire
0	0	`0000`	8	8	`1000`
1	1	`0001`	9	9	`1001`
2	2	`0010`	A	10	`1010`
3	3	`0011`	B	11	`1011`
4	4	`0100`	C	12	`1100`
5	5	`0101`	D	13	`1101`
6	6	`0110`	E	14	`1110`
7	7	`0111`	F	15	`1111`

de la mémoire ou d'un fichier. On le reconnaît à des préfixes comme `0x` (`0xE9`, dans les langages de programmation) ou `U+` (`U+00E9`, pour les caractères Unicode, section 6.4). On le retrouve dans les adresses matérielles des cartes réseau (`3C:22:FB:9A:01:7E`), dans les adresses Internet de nouvelle génération (IPv6) et, surtout, dans les codes de couleur du Web.

6.3.4 Exemple phare : les codes de couleur du Web

Un écran forme chaque couleur en mélangeant trois lumières : rouge, verte et bleue. C'est une synthèse *additive* : rouge et vert donnent du jaune, les trois lumières à pleine intensité donnent du blanc, et l'absence de lumière donne du noir. On le constate à la loupe sur un écran, où chaque pixel est formé de trois petites sources lumineuses.

Définition 6.3 **Codage RVB**

Le **codage RVB** (*RGB* en anglais) décrit une couleur par un triplet (R, V, B) d'entiers compris entre 0 et 255, qui donnent l'intensité des lumières rouge, verte et bleue. Chaque composante occupe donc un octet, et une couleur, 3 octets ou 24 bits.

Il y a ainsi $256^3 = 2^{24} = 16\,777\,216$ couleurs différentes, soit un peu plus de 16 millions. En HTML et en CSS, les langages des pages Web, on écrit une couleur sous la forme `#RRVVBB` : un dièse, suivi des trois octets écrits en hexadécimal. Ainsi, `#FF0000` est le rouge pur $(255, 0, 0)$, `#FFFF00` est le jaune $(255, 255, 0)$, et `#808080` est un gris moyen $(128, 128, 128)$. La couleur des titres de ce livre est `#1C3F6E`, c'est-à-dire

$$(1\mathrm{C}, 3\mathrm{F}, 6\mathrm{E})_{16} = (1 \times 16 + 12,\ 3 \times 16 + 15,\ 6 \times 16 + 14) = (28, 63, 110)\,:$$

un bleu foncé, où le bleu domine, avec un peu de vert et encore moins de rouge.

Ce codage permet de « lire » une couleur sans l'afficher. Le premier chiffre de chaque paire est le plus important, puisqu'il vaut 16 fois plus que le second : `#E0E0E0` est un gris très pâle, `#202020` un gris très foncé. Une couleur dont les trois paires sont égales est un gris ; il y en a 256, du noir `#000000` au blanc `#FFFFFF`. La couleur « complémentaire » s'obtient en soustrayant chaque composante de `FF` : le complémentaire du rouge `#FF0000` est le cyan `#00FFFF`. Enfin, lorsque chaque paire est formée de deux chiffres identiques, on peut abréger : `#F80` signifie `#FF8800`.

Outil 6.1 **Explorer les codes de couleur**

Dans un moteur de recherche, taper « color picker » affiche un sélecteur de couleur qui donne simultanément les codes hexadécimal et RVB. Les outils de développe-



FIGURE 6.3 : Quelques couleurs et leur code hexadécimal. Les trois paires de chiffres donnent, dans l'ordre, l'intensité du rouge, du vert et du bleu, de `00` (éteint) à `FF` (pleine intensité). Les gris sont les couleurs dont les trois composantes sont égales.

ment des navigateurs (touche F12) montrent les couleurs de n'importe quelle page Web. Dans GeoGebra, la commande `SetColor(objet, "#1C3F6E")` colore un objet à partir d'un code hexadécimal.
Enfin, Python convertit d'une base à l'autre en une ligne. Dans un sens, `int("6E", 16)` lit la chaîne `"6E"` comme un nombre écrit en base seize, et donne `110`. Dans l'autre sens, la fonction `format` transforme un nombre en chaîne de caractères selon un *code de format*. Dans `format(110, "02X")`, ce code se lit de droite à gauche :

- `X` demande l'écriture hexadécimale, avec des lettres majuscules (`x` donnerait des minuscules, `o` l'octal et `b` le binaire) ;
- `2` demande au moins deux caractères ;
- `0` demande de compléter au besoin par des zéros à gauche.

Ainsi, `format(110, "02X")` donne `'6E'`, et `format(10, "02X")` donne `'0A'` plutôt que `'A'` : c'est ce qu'il faut pour un code de couleur, où chaque composante doit occuper exactement deux chiffres. Le code de `#1C3F6E` s'obtient donc avec `"#" + format(28, "02X") + format(63, "02X") + format(110, "02X")`. De même, `format(2026, "b")` donne `'11111101010'`, et `format(5, "08b")` donne `'00000101'`, l'écriture de 5 sur un octet.

Pour la classe 6.2 **Un défi pour la classe**

Les codes de couleur se prêtent à une activité courte et motivante. On projette des carrés colorés, et les élèves doivent deviner leur code hexadécimal (ou, à l'inverse, prédire la couleur d'un code donné), puis vérifier avec un sélecteur de couleur. Les élèves découvrent la synthèse additive (pourquoi `#FFFF00` est-il jaune ?), la valeur de position (pourquoi `#80` est-il « à moitié » et `#08` « presque rien » ?) et le dénombrement ($16^6 = 256^3$). C'est aussi une occasion de montrer que la notation positionnelle n'est pas liée à la base dix.

6.4 Encoder du texte

6.4.1 Un encodage, c'est une convention

Un ordinateur ne connaît pas les lettres. Pour stocker un texte, il faut une *convention*, partagée par celui qui écrit et celui qui lit, qui associe un nombre à chaque caractère.

Définition 6.4 — Jeu de caractères, encodage

Un **jeu de caractères** associe à chaque caractère d'un répertoire (lettres, chiffres, ponctuation, symboles) un entier naturel appelé son **point de code**. Un **encodage** précise ensuite comment chaque point de code est représenté par une suite d'octets.

Pour les encodages anciens, les deux notions se confondent : chaque caractère est représenté par un seul octet, qui est l'écriture binaire de son point de code. Nous verrons que ce n'est plus le cas avec l'UTF-8. Un fichier texte n'indique généralement pas l'encodage utilisé : le logiciel qui le lit doit le deviner ou le connaître. C'est la source de bien des problèmes.

6.4.2 Du télégraphe à l'ASCII

Les encodages sont plus anciens que l'ordinateur. Le code Morse (vers 1840) représente déjà les lettres par des suites de deux symboles, le point et le trait, mais de longueurs variables. Le code inventé en 1874 par l'ingénieur français Émile Baudot pour le télégraphe utilise 5 bits par caractère, soit $2^5 = 32$ combinaisons. C'est trop peu pour les lettres, les chiffres et la ponctuation, et Baudot utilise une astuce : deux combinaisons spéciales font basculer entre un mode « lettres » et un mode « chiffres », un peu comme la touche de verrouillage des majuscules. L'unité de vitesse de transmission, le *baud*, porte son nom.

En 1963, l'organisme de normalisation américain publie l'ASCII (*American Standard Code for Information Interchange*), qui utilise 7 bits, soit 128 caractères numérotés de 0 à 127. Les 32 premiers sont des caractères de contrôle, invisibles, hérités des téléscripteurs : 9 est la tabulation, 10 le saut de ligne, 13 le retour de chariot. Viennent ensuite l'espace (32), la ponctuation, les chiffres, les majuscules et les minuscules.

Table 6.2 : Un extrait de la table ASCII. Les points de code sont donnés en décimal, en hexadécimal et en binaire.

Caractère	Décimal	Hexa	Binaire
espace	32	`20`	`0100000`
`0`	48	`30`	`0110000`
`9`	57	`39`	`0111001`
`A`	65	`41`	`1000001`
`Z`	90	`5A`	`1011010`
`a`	97	`61`	`1100001`
`z`	122	`7A`	`1111010`

La table n'est pas arbitraire, et l'écriture binaire en révèle la logique. Les chiffres `0` à `9` occupent les points de code `0x30` à `0x39` : la valeur d'un chiffre se lit dans ses 4 derniers bits. Une majuscule et la minuscule correspondante diffèrent de $32 = 2^5$: `A` est `1000001` et `a` est `1100001`. Passer de la majuscule à la minuscule revient donc à changer *un seul bit*, une opération que les circuits effectuent instantanément.

6.4.3 Les accents : un octet pour tout le monde ?

L'ASCII suffit pour l'anglais, mais pas pour le français, qui a besoin de « é », « è », « à », « ç », « ô », « œ »… Les ordinateurs manipulant des octets de 8 bits, il restait 128 points de code libres, de 128 à 255. Chaque région du monde les a remplis à sa façon. La norme ISO 8859-1 (ou *Latin-1*, 1987) couvre les langues d'Europe de l'Ouest ; Microsoft en a fait une variante, Windows-1252, qui utilise aussi les points de code 128 à 159 pour y placer notamment les guillemets typographiques, le symbole de l'euro (`0x80`) et les ligatures « œ » (`0x9C`) et « Œ » (`0x8C`), absentes de Latin-1. Le bloc-notes de Windows a longtemps appelé cet encodage « ANSI », du nom de l'organisme de normalisation américain, une appellation impropre puisque l'ANSI ne l'a jamais normalisé (figure 6.4). Dans tous ces encodages, « é » porte le numéro 233, soit `0xE9` ou `11101001`.



Figure 6.4 : Au moment d'enregistrer un fichier, le bloc-notes de Windows propose de choisir son encodage. Ici, « ANSI » désigne en fait Windows-1252. Les versions récentes enregistrent par défaut en UTF-8.

Le problème est qu'il existait des dizaines de tables incompatibles : le point de code 233 désignait « é » en Europe de l'Ouest, une lettre cyrillique dans une autre table, et une lettre grecque ou hébraïque ailleurs. Un courriel écrit à Montréal pouvait devenir illisible à Moscou. Et l'approche était sans issue pour le chinois ou le japonais, qui comptent des milliers de caractères.

6.4.4 Unicode : un numéro pour chaque caractère

Le projet Unicode, lancé à la fin des années 1980, adopte un point de vue radicalement différent : attribuer un numéro unique à *chaque* caractère de *chaque* système d'écriture, passé ou présent. La version 1.0 paraît en 1991. L'espace des points de code va de `U+0000` à `U+10FFFF` (on les écrit en hexadécimal, précédés de `U+`), soit 1 114 112 positions ; la version 17.0, publiée en 2025, en attribue près de 160 000 [The25]. On y trouve les alphabets latin, grec, cyrillique, arabe, les idéogrammes chinois, le syllabaire inuktitut, les symboles mathématiques comme `U+222B` (∫) et les émojis. Les 256 premiers points de code reprennent Latin-1 : « é » reste `U+00E9`, soit 233. Mais « € » devient `U+20AC`, soit 8364, et le « visage souriant » des émojis est `U+1F600`, soit 128 512.

Unicode est un jeu de caractères, pas un encodage : il dit quel *numéro* porte chaque caractère, mais pas comment écrire ce numéro en octets. La solution la plus simple, utiliser 4 octets par caractère (l'encodage UTF-32), quadruple la taille des textes anglais et rend illisibles tous les fichiers ASCII existants.

6.4.5 L'UTF-8

La solution qui s'est imposée a été imaginée en septembre 1992 par Ken Thompson et Rob Pike, deux informaticiens des laboratoires Bell, esquissée, selon Pike, sur un napperon dans un restaurant du New Jersey [Pik03]. C'est un encodage à *longueur variable* : un caractère occupe de 1 à 4 octets, selon la taille de son point de code.

Définition 6.5 **L'encodage UTF-8**

L'**UTF-8** représente un point de code p par 1, 2, 3 ou 4 octets, selon le gabarit suivant, où les bits notés x sont ceux de l'écriture binaire de p, complétée par des 0 à gauche :

Points de code	Bits	Octets
`U+0000` à `U+007F`	7	`0xxxxxxx`
`U+0080` à `U+07FF`	11	`110xxxxx 10xxxxxx`
`U+0800` à `U+FFFF`	16	`1110xxxx 10xxxxxx 10xxxxxx`
`U+10000` à `U+10FFFF`	21	`11110xxx 10xxxxxx 10xxxxxx 10xxxxxx`

Le premier octet annonce la longueur : il commence par `0` pour un caractère d'un seul octet, par `110` pour deux octets, `1110` pour trois et `11110` pour quatre. Les octets suivants, dits de *continuation*, commencent tous par `10` et transportent 6 bits chacun.

Exemple 6.3 **Encoder « é », « € » et un émoji**

La lettre « é ». Son point de code est $233 = (11101001)_2$, compris entre `0x80` et `0x7FF` : il faut deux octets et 11 bits. On écrit 233 sur 11 bits, `00011101001`, que l'on répartit en $5 + 6$ bits :

$$\underbrace{00011}_{5}\ \underbrace{101001}_{6} \longmapsto \texttt{110 00011 10 101001} = \texttt{C3 A9}.$$

En UTF-8, « é » occupe donc deux octets, `C3 A9`, alors qu'il n'en occupait qu'un, `E9`, en Windows-1252.
Le symbole « € ». Son point de code `U+20AC` exige trois octets. Ses 16 bits, `0010 0000 1010 1100`, se répartissent en $4 + 6 + 6$:

$$\underbrace{0010}_{4}\ \underbrace{000010}_{6}\ \underbrace{101100}_{6} \longmapsto \texttt{1110 0010 10 000010 10 101100} = \texttt{E2 82 AC}.$$

Le visage souriant `U+1F600`. Il faut quatre octets. Sur 21 bits, $1F600_{16}$ s'écrit `000 011111 011000 000000`, d'où `11110000 10011111 10011000 10000000`, soit `F0 9F 98 80`.

Ce gabarit, qui peut sembler compliqué, a des propriétés remarquables.

1. *Compatibilité avec l'ASCII.* Les 128 caractères ASCII s'encodent sur un octet, exactement comme en ASCII. Tout fichier ASCII est donc, sans modification, un fichier UTF-8 valide, et un texte anglais n'occupe pas plus de place qu'avant.
2. *Pas de confusion.* Les octets d'un caractère de plusieurs octets sont tous supérieurs ou égaux à 128. Un tel octet ne peut donc jamais être pris pour un caractère ASCII, comme la barre oblique `/` d'un chemin de fichier ou le point d'un nom de domaine.
3. *Autosynchronisation.* En lisant n'importe quel octet, on sait s'il commence un caractère (`0` ou `11`) ou s'il en continue un (`10`). Si un octet est perdu ou endommagé, un seul caractère est affecté, et la lecture reprend correctement au caractère suivant.
4. *Ordre préservé.* Trier des chaînes UTF-8 octet par octet donne le même ordre que trier les points de code.

Ces qualités expliquent son succès : en 2025, plus de 98 % des pages Web sont encodées en UTF-8, et c'est l'encodage par défaut de Python, de macOS, de Linux et des versions récentes de Windows.

Remarque 6.2 **Pourquoi « Ã© » au lieu de « é » ?**

Tout le monde a déjà reçu un courriel ou ouvert un fichier où « é » apparaissait sous la forme « Ã© ». Ce phénomène, que les Japonais appellent *mojibake* (« caractères transformés »), a une explication simple. Le texte a été écrit en UTF-8, où « é » est `C3 A9`, puis lu comme s'il était encodé en Windows-1252, où chaque octet est un caractère : `C3` est « Ã » et `A9` est « © ». De même, « ç » (`C3 A7`) devient « Ã§ », et « € » (`E2 82 AC`) devient « â,¬ ». L'erreur inverse, lire comme de l'UTF-8 un texte écrit en Windows-1252, produit plutôt un caractère de remplacement (souvent un losange contenant un point d'interrogation), puisque l'octet `E9` isolé n'est pas une suite UTF-8 valide. Reconnaître ces symptômes permet de diagnostiquer immédiatement un problème d'encodage.

Pour la classe 6.3 **Des messages secrets en binaire**

Faire décoder un court message écrit en binaire, octet par octet, à l'aide de la table ASCII est une activité classique, qui fait travailler la conversion de base. Le passage à l'UTF-8 en fait un vrai problème : pourquoi le message « Ça va ? » contient-il 7 octets et non 6 ? On peut aussi faire fabriquer aux élèves des bracelets de perles de deux couleurs qui « écrivent » leurs initiales en binaire.

6.5 Mesurer l'information

6.5.1 Le bit et l'octet

Le bit est la plus petite quantité d'information : la réponse à une question par oui ou par non. Les ordinateurs ne manipulent cependant pas les bits un par un, mais par groupes.

Définition 6.6 **Octet**

Un **octet** est une chaîne de 8 bits. Il peut prendre $2^8 = 256$ valeurs, et s'écrit avec deux chiffres hexadécimaux.

Le mot anglais correspondant est *byte*, et non *bit* : c'est un véritable faux ami. Le terme *byte* a été créé en 1956 par Werner Buchholz, chez IBM, pour désigner le groupe de bits qui code un caractère ; sa taille a d'abord varié d'une machine à l'autre, avant de se fixer à 8 bits dans les années 1960. Le français « octet » a l'avantage de ne laisser aucun doute. Les abréviations sont à surveiller : `o` ou `B` pour l'octet, `b` ou `bit` pour le bit. Les fournisseurs d'accès Internet expriment les débits en *bits* par seconde : un forfait de 100 Mbit/s permet de télécharger au mieux $100 \div 8 = 12{,}5$ mégaoctets par seconde.

La mémoire d'un ordinateur est organisée comme un immense tableau d'octets numérotés : chaque octet a une **adresse**, un entier que l'on écrit habituellement en hexadécimal. Le nombre de bits utilisés pour écrire les adresses limite la taille de la mémoire : avec des adresses de 32 bits, on ne peut désigner que 2^{32} octets, soit 4 Gio (voir plus bas). C'est pourquoi les versions 32 bits de Windows ne pouvaient pas utiliser plus de 4 Gio de mémoire vive, une limite levée par le passage aux processeurs 64 bits.

6.5.2 Préfixes décimaux et préfixes binaires

Pour les grandes quantités, on utilise les préfixes du Système international d'unités (SI) : kilo, méga, giga, téra, péta, qui désignent des puissances de 1000.

Table 6.3 : Les multiples décimaux de l'octet, avec des ordres de grandeur.

Nom	Symbole	Nombre d'octets	Ordre de grandeur
kilooctet	ko	10^3	une page de texte brut : quelques ko
mégaoctet	Mo	10^6	une photo de téléphone : quelques Mo
gigaoctet	Go	10^9	un film en haute définition : quelques Go
téraoctet	To	10^{12}	un disque dur actuel : quelques To
pétaoctet	Po	10^{15}	les archives d'une grande bibliothèque numérique
exaoctet	Eo	10^{18}	le trafic Internet mondial en quelques jours

Or, l'informatique a une préférence naturelle pour les puissances de 2, et il se trouve que $2^{10} = 1024$ est très proche de 1000. Pendant des décennies, les informaticiens ont donc employé « kilooctet » pour 1024 octets, « mégaoctet » pour 1024^2 octets, et ainsi de suite. L'écart, de 2,4 % pour le kilo, grandit avec l'exposant. Pour lever l'ambiguïté, la Commission électrotechnique internationale (CEI) a introduit en 1998 des **préfixes binaires** : kibi, mébi, gibi, tébi, formés des deux premières lettres du préfixe décimal suivies de « bi », pour *binaire*.

Table 6.4 : Les multiples binaires de l'octet et leur écart avec les multiples décimaux correspondants.

Nom	Symbole	Nombre d'octets	Écart
kibioctet	Kio	$2^{10} = 1024$	2,4 % de plus que 1 ko
mébioctet	Mio	$2^{20} = 1\,048\,576$	4,9 % de plus que 1 Mo
gibioctet	Gio	$2^{30} = 1\,073\,741\,824$	7,4 % de plus que 1 Go
tébioctet	Tio	$2^{40} = 1\,099\,511\,627\,776$	10,0 % de plus que 1 To

6.5.3 Où sont passés les gigaoctets ?

Les fabricants de disques durs ont toujours utilisé les préfixes décimaux, conformément au SI. Un disque de « 200 Go » contient bien 200 milliards d'octets. Mais Windows calcule la taille en gibioctets tout en affichant le symbole « Go ». Sur la figure 6.5, le disque de 200 048 447 488 octets est affiché comme un disque de 186 « Go », car

$$\frac{200\,048\,447\,488}{2^{30}} \approx 186{,}3.$$

Personne ne ment sur la quantité d'octets : c'est l'unité qui change en chemin. Le système macOS, lui, affiche les tailles en unités décimales depuis 2009. Un disque de 1 To apparaît dans Windows comme un disque de 931 « Go », soit 7 % de moins.

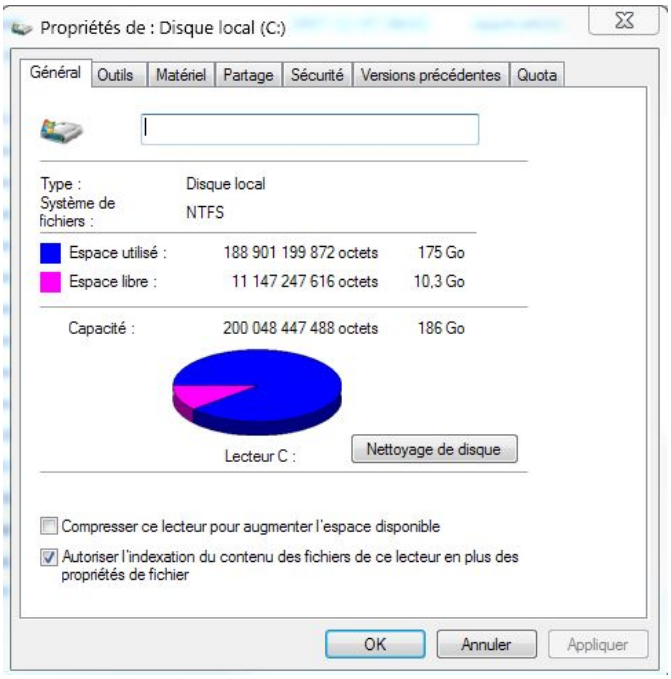


FIGURE 6.5 : Les propriétés d'un disque dur sous Windows. La capacité de 200 048 447 488 octets est affichée comme 186 « Go » : il s'agit en fait de gibioctets.

Remarque 6.3 **La disquette de « 1,44 Mo » et le CD de « 700 Mo »**

Revenons à la figure 6.1. La disquette de 3,5 pouces dite « haute densité » contient 1 474 560 octets. Ce n'est ni 1,44 million d'octets, ni 1,44 Mio, mais 1440×1024 octets : le « Mo » de la disquette est une unité hybride de 1000×1024 octets ! Un CD-R de 80 minutes contient 360 000 secteurs de 2048 octets de données, soit 737 280 000 octets, ou environ 703 Mio : son « Mo » est un mébioctet. Le rapport est exactement

$$\frac{737\,280\,000}{1\,474\,560} = 500 \text{ disquettes.}$$

Pourquoi parle-t-on de 80 minutes ? Sur un CD audio, chaque secteur contient 2352 octets de son, au lieu de 2048 octets de données : un CD-ROM réserve les 304 octets restants à une correction d'erreurs supplémentaire, une erreur dans un programme étant plus grave qu'un petit craquement dans une chanson. Le son d'un CD est échantillonné 44 100 fois par seconde, sur 16 bits (2 octets), en stéréo (2 canaux), soit 176 400 octets par seconde. On vérifie que

$$\frac{360\,000 \times 2352}{176\,400} = 4800 \text{ secondes} = 80 \text{ minutes.}$$

6.6 Mémoire vive et mémoire de stockage

6.6.1 Le bureau et la bibliothèque

Un ordinateur possède deux grandes sortes de mémoire, que l'on confond souvent parce qu'elles se mesurent avec les mêmes unités. Une analogie aide à les distinguer. La **mémoire de stockage** (disque dur ou SSD) est la bibliothèque : elle est vaste, peu coûteuse, et

ce qu'on y range y reste, même lorsque l'ordinateur est éteint. La **mémoire vive** (ou RAM, pour *Random Access Memory*) est la surface du bureau : elle est plus petite et plus chère, mais tout ce qui s'y trouve est à portée de main. Pour travailler sur un document, le processeur le copie de la bibliothèque vers le bureau ; c'est ce que fait l'ordinateur quand on « ouvre » un fichier. À l'extinction, le bureau est vidé : la mémoire vive est *volatile*, et c'est pourquoi un document non enregistré est perdu lors d'une panne de courant.

Table 6.5 : Ordres de grandeur pour un ordinateur personnel en 2025. Les temps d'accès sont des valeurs typiques ; la dernière colonne les ramène à une échelle humaine, en supposant qu'un accès à la mémoire vive dure une seconde.

Mémoire	Capacité typique	Temps d'accès	À l'échelle humaine
mémoire cache du processeur	quelques Mo	1 à 10 ns	moins d'une seconde
mémoire vive (DRAM)	8 à 64 Go	≈ 100 ns	1 seconde
disque SSD	0,5 à 4 To	≈ 100 µs	17 minutes
disque dur magnétique	1 à 30 To	≈ 10 ms	plus d'une journée

Le tableau 6.5 montre que l'écart de vitesse est considérable : un disque dur magnétique est environ 100 000 fois plus lent que la mémoire vive. Si la mémoire vive répondait en une seconde, le disque dur mettrait plus d'une journée. C'est ce qui explique qu'un ordinateur dont la mémoire vive est saturée, et qui doit utiliser le disque comme mémoire d'appoint, devienne si lent.

6.6.2 Comment un bit est-il inscrit ?

Revenons maintenant à la question posée au début du chapitre : où se trouvent physiquement les 0 et les 1 ? La réponse dépend du support, mais le principe est toujours le même : un phénomène physique qui peut se trouver dans deux états facilement distinguables et suffisamment stables.

La mémoire vive (DRAM). Chaque bit est stocké par un minuscule condensateur, chargé ou déchargé, associé à un transistor qui sert de porte d'accès. Un condensateur chargé se vide lentement par fuites : la mémoire doit donc relire et réécrire chacun de ses bits toutes les 64 millisecondes environ. C'est le *rafraîchissement*, qui explique le « D », pour *dynamique*, de DRAM, et le caractère volatile de cette mémoire. Une barrette de 16 Go contient ainsi environ 137 milliards de condensateurs.

Le disque dur magnétique. Des plateaux de verre ou d'aluminium, recouverts d'une couche magnétique, tournent à 5400 ou 7200 tours par minute. Une tête de lecture, qui vole à quelques nanomètres de la surface, aimante de minuscules régions dans un sens ou dans l'autre, puis détecte ces orientations. En pratique, les bits sont codés par les *changements* d'orientation plutôt que par les orientations elles-mêmes, ce qui facilite la synchronisation de la lecture. Le temps d'accès de quelques millisecondes est celui qu'il faut pour déplacer la tête et attendre que le bon secteur passe sous elle.

Le disque optique (CD, DVD, Blu-ray). La surface porte une spirale de minuscules creux. Un laser la parcourt, et la lumière réfléchie change d'intensité à chaque début et à chaque fin de creux ; ces transitions codent les bits. Le DVD et le Blu-ray stockent davantage parce que leurs lasers, de longueur d'onde plus courte (rouge, puis bleu-violet), permettent des creux plus petits et plus rapprochés.

La mémoire flash (clés USB, cartes mémoire, SSD). Chaque bit est stocké dans un transistor d'un type particulier, à *grille flottante* (figure 6.6). Dans un transistor ordinaire, la tension appliquée sur une électrode, la *grille de commande*, décide si le courant peut passer ou

non entre deux autres électrodes, la *source* et le *drain* : le transistor est un interrupteur commandé électriquement. Le transistor à grille flottante possède en plus une seconde grille, coincée entre la grille de commande et le silicium, et complètement entourée d'un isolant. Aucun fil ne la relie au reste du circuit, d'où son nom.

— *Écrire un 0.* On applique une tension élevée (une quinzaine de volts) sur la grille de commande. Des électrons du silicium traversent alors la très mince couche isolante, épaisse de quelques nanomètres, par un phénomène quantique appelé l'*effet tunnel*, et se retrouvent piégés dans la grille flottante.

— *Conserver.* Une fois la tension coupée, l'isolant empêche les électrons de repartir. Ils restent piégés pendant des années, même sans alimentation : la mémoire flash est *non volatile*.

— *Lire.* On applique sur la grille de commande une tension modérée, bien choisie. Si la grille flottante est vide, cette tension suffit à faire conduire le transistor : le courant passe, et l'on lit un 1. Si elle contient des électrons, leur charge négative fait écran, le transistor ne conduit pas, et l'on lit un 0.

— *Effacer.* Une tension élevée de sens opposé fait ressortir les électrons. Pour simplifier les circuits, on efface d'un seul coup un bloc entier de cellules, souvent plusieurs centaines de kilooctets : c'est ce « coup de flash » qui a donné son nom à cette mémoire, mise au point chez Toshiba dans les années 1980.

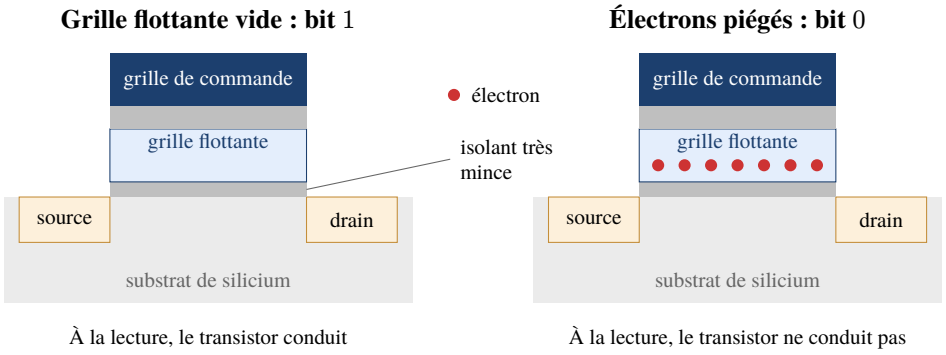


FIGURE 6.6 : Coupe schématique d'une cellule de mémoire flash. À gauche, la grille flottante est vide : à la lecture, le transistor conduit, et la cellule contient un 1. À droite, des électrons ont été piégés dans la grille flottante par effet tunnel ; ils empêchent le transistor de conduire, et la cellule contient un 0.

Chaque passage d'électrons à travers l'isolant l'abîme un peu. Une cellule ne supporte donc qu'un nombre limité d'effacements, de quelques milliers à une centaine de milliers selon la technologie. Pour qu'une clé USB ou un SSD ne s'use pas prématurément, son contrôleur répartit les écritures sur l'ensemble des cellules, plutôt que de réécrire toujours aux mêmes endroits.

Sur tous ces supports, des erreurs se produisent : un bit de mémoire vive peut être inversé par une particule cosmique, un creux de CD peut être rayé. Les données sont donc toujours accompagnées de bits supplémentaires qui permettent de détecter et de corriger les erreurs. Nous avons rencontré l'idée de base de ces *codes correcteurs* au chapitre 2, avec la distance de Hamming (définition 2.2).

6.7 Un historique des capacités de stockage

Le premier disque dur, l'IBM 350, livré en 1956 avec l'ordinateur RAMAC 305, contenait 5 millions de caractères de 6 bits, soit environ 3,75 Mo. Il se composait de 50 plateaux de 24 pouces (61 cm) de diamètre, pesait environ une tonne et était loué, avec l'ordinateur, quelques milliers de dollars par mois. En 1973, l'IBM 3340, surnommé « Winchester », introduit la tête de lecture qui « vole » au-dessus du plateau, principe encore utilisé aujourd'hui. Le premier disque dur de 5,25 pouces destiné aux micro-ordinateurs, le Seagate ST-506 (1980), contenait 5 Mo et coûtait environ 1500 dollars américains. Le gigaoctet est atteint en 1980 par les gros systèmes (IBM 3380) et en 1991 par les disques de 3,5 pouces ; le téraoctet, en 2007. En 2024, les premiers disques utilisant l'enregistrement magnétique assisté par la chaleur (HAMR), où un laser chauffe brièvement la surface avant l'écriture, dépassent les 30 To.

Table 6.6 : Quelques jalons de l'histoire du disque dur. Les capacités sont celles des modèles les plus grands de leur catégorie ; les valeurs sont arrondies.

Année	Modèle	Capacité	Remarque
1956	IBM 350 (RAMAC)	3,75 Mo	50 plateaux de 61 cm, environ une tonne
1973	IBM 3340	70 Mo	tête volante « Winchester »
1980	Seagate ST-506	5 Mo	premier disque de 5,25 po
1980	IBM 3380	2,52 Go	premier disque de plus d'un gigaoctet
1991	IBM 0663	1 Go	3,5 po, tête magnétorésistive
1997	IBM Deskstar 16GP	16,8 Go	tête à magnétorésistance géante
2007	Hitachi Deskstar 7K1000	1 To	premier disque d'un téraoctet
2013	HGST Ultrastar He6	6 To	boîtier rempli d'hélium
2024	Seagate Exos M	30 To	enregistrement assisté par la chaleur

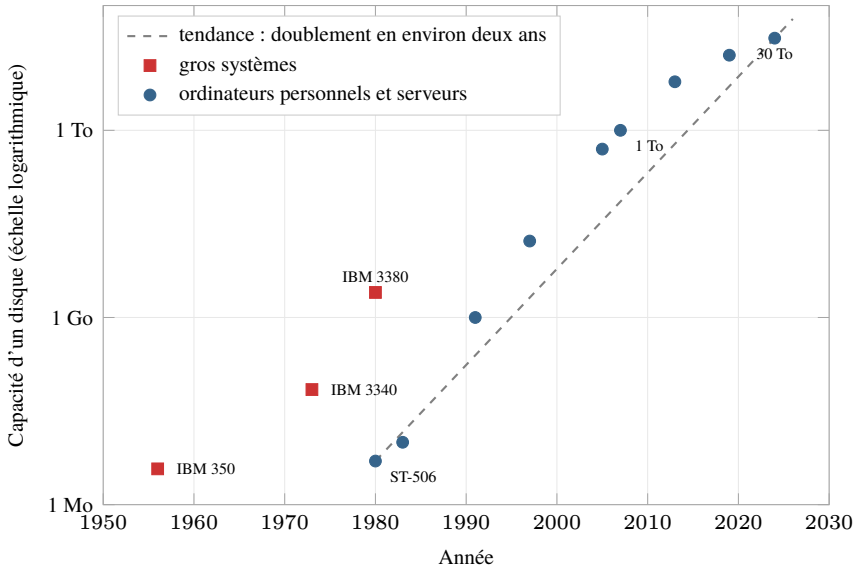


Figure 6.7 : La capacité des disques durs, sur une échelle logarithmique. Les points ronds suivent approximativement une droite : la croissance est exponentielle. Entre le ST-506 de 1980 (5 Mo) et les disques de 30 To de 2024, la capacité a été multipliée par environ 6 millions, ce qui correspond à un doublement tous les deux ans en moyenne.

Sur une échelle logarithmique, une croissance exponentielle se traduit par une droite. La figure 6.7 montre que la capacité des disques a suivi une telle tendance pendant plus de

quarante ans. Faisons le calcul : de 5 Mo en 1980 à 30 To en 2024, le facteur est

$$\frac{30 \times 10^{12}}{5 \times 10^{6}} = 6 \times 10^{6} \approx 2^{22{,}5},$$

soit environ 22,5 doublements en 44 ans, un doublement tous les deux ans. Pendant ce temps, le prix du gigaoctet est passé d'environ 300 000 dollars (1500 dollars pour 5 Mo) à environ deux cents, une division par plus de dix millions. Cette tendance, souvent appelée « loi de Kryder », est l'équivalent pour le stockage de la loi de Moore pour les transistors. Elle s'est toutefois nettement ralentie depuis 2010, à mesure que l'on s'approche des limites physiques de l'enregistrement magnétique. Dans les ordinateurs personnels, le disque dur magnétique a d'ailleurs été largement remplacé par le SSD, plus rapide et plus robuste ; il reste cependant le support le moins coûteux pour les grands centres de données.

Les supports amovibles racontent la même histoire. La disquette de 8 pouces (1971) contenait environ 80 ko ; celle de 3,5 pouces, 1,44 « Mo ». Le CD-ROM (1985) contenait 650 à 700 Mo, le DVD (1996), 4,7 Go par couche, et le Blu-ray (2006), 25 Go par couche. Les premières clés USB, vers 2000, contenaient 8 Mo ; on en trouve aujourd'hui de plusieurs téraoctets, dans un volume de quelques centimètres cubes.

Pour la classe 6.4 **Une fonction exponentielle en contexte**

Les données du tableau 6.6 constituent un bon contexte pour l'étude des fonctions exponentielles et logarithmiques en cinquième secondaire. Les élèves peuvent représenter les données sur une échelle ordinaire, constater que le graphique est illisible, passer au logarithme de la capacité, ajuster une droite, puis en déduire le temps de doublement. La comparaison avec la loi de Moore, et la discussion de la validité d'une extrapolation (la tendance va-t-elle se poursuivre ?), complètent naturellement l'activité.

6.8 Les images numériques

On appelle **image numérique** toute image acquise, créée, traitée ou stockée sous forme binaire. Une photo prise avec un téléphone, une capture d'écran, un logo dessiné à l'ordinateur ou une figure de ce livre sont des images numériques. On en distingue deux grandes familles, qui reposent sur deux idées très différentes : décrire l'image *point par point*, ou la décrire *par des objets géométriques*.

6.8.1 Les images matricielles

Définition 6.7 **Image matricielle**

Une **image matricielle** est décrite par un tableau rectangulaire (une matrice) dont chaque case, appelée **pixel** (de l'anglais *picture element*), contient une couleur. La **définition** de l'image est le nombre de pixels, souvent donné sous la forme « largeur × hauteur ».

La taille des données d'une image matricielle dépend du nombre de bits utilisés pour décrire la couleur de chaque pixel, ce qu'on appelle la **profondeur** de l'image :

— pour une image en noir et blanc, 1 bit par pixel suffit ;



FIGURE 6.8 : En agrandissant une image matricielle, on finit par voir les pixels, des carrés de couleur uniforme. Source : Ivanoff, « DigitalPicture.jpg », *Wikimedia Commons*.

— pour une image en 256 tons de gris, il faut 8 bits, soit 1 octet par pixel ;
— pour une image en couleurs RVB (définition 6.3), il faut 3 octets par pixel ; on ajoute parfois un quatrième octet, la composante *alpha*, qui décrit la transparence.

Exemple 6.4 **La taille d'une photo**

Un téléphone prend des photos de 4000×3000 pixels, soit 12 millions de pixels (« 12 mégapixels »). En couleurs RVB, sans compression, les données occupent

$$4000 \times 3000 \times 3 = 36\,000\,000 \text{ octets},$$

soit 36 Mo. Pourtant, le fichier JPEG produit par le téléphone n'en occupe généralement que 3 à 5. La différence est due à la compression, que nous décrirons plus bas.

Un fichier d'image matricielle comprend un **en-tête**, qui précise notamment les dimensions de l'image et la façon dont les données sont encodées, suivi des **données** elles-mêmes. L'en-tête peut aussi contenir des **métadonnées** : la date, le modèle de l'appareil et, souvent, les coordonnées GPS du lieu de la prise de vue. Les formats les plus connus sont BMP (*bitmap*, sans compression), PNG (*Portable Network Graphics*, avec compression sans perte) et JPEG (*Joint Photographic Experts Group*, avec compression avec perte).

Exemple 6.5 **Une image écrite à la main**

La famille de formats Netpbm, créée dans les années 1980, permet d'écrire une image dans un simple fichier texte. Le fichier suivant, que l'on peut taper dans un bloc-notes et enregistrer sous le nom `visage.pbm`, décrit une image en noir et blanc de 10×9 pixels. La première ligne, `P1`, est l'en-tête qui annonce le format ; la deuxième donne la largeur et la hauteur ; viennent ensuite les pixels, ligne par ligne, 1 pour noir et 0 pour blanc.

```
P1
10 9
0 0 0 0 0 0 0 0 0 0
0 0 1 1 0 0 1 1 0 0
0 0 1 1 0 0 1 1 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 1 0 0 0 0 0 0 1 0
0 0 1 0 0 0 0 1 0 0
0 0 0 1 1 1 1 0 0 0
0 0 0 0 0 0 0 0 0 0
```

Le logiciel libre GIMP, entre autres, ouvre ce fichier et affiche l'image de la figure 6.9. Le format `P3` fonctionne de la même façon en couleurs : chaque pixel y est décrit par trois entiers de 0 à 255, ses composantes RVB (exercice 6.19).

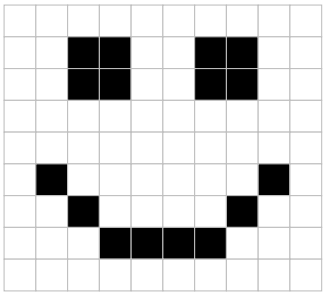


FIGURE 6.9 : L'image décrite par le fichier `visage.pbm` de l'exemple 6.5, agrandie : chaque case est un pixel.

6.8.2 La compression

Compresser une image, c'est la décrire avec moins de bits que la description pixel par pixel. On distingue deux sortes de compression.

La **compression sans perte** permet de retrouver l'image d'origine exactement. Elle exploite les régularités des données. L'idée la plus simple, le *codage par plages* (*run-length encoding*), remplace une suite de pixels identiques par leur nombre. La ligne

`0000000000000111111000000000`

de 28 pixels peut s'écrire « 13 blancs, 6 noirs, 9 blancs ». Cette méthode est très efficace pour un logo ou une capture d'écran, qui comportent de grandes plages de couleur uniforme, et beaucoup moins pour une photo, où deux pixels voisins sont rarement identiques. Le format PNG utilise des méthodes plus raffinées, mais fondées sur la même idée : repérer et abréger les répétitions.

La **compression avec perte** accepte de modifier légèrement l'image, en éliminant des détails que l'œil remarque peu. Le format JPEG découpe l'image en blocs de 8×8 pixels

et décrit chaque bloc comme une combinaison de motifs plus ou moins fins, à l'aide d'une transformation mathématique voisine des séries de Fourier. L'œil étant peu sensible aux variations très fines, on conserve avec précision les coefficients des motifs grossiers, et on arrondit fortement, voire on élimine, ceux des motifs fins. Un bloc de couleur presque uniforme se décrit alors avec très peu d'information, alors qu'un bloc qui contient un contour net en exige davantage. On obtient couramment un rapport de compression de 10 pour 1 sans perte de qualité visible ; à des taux plus élevés, les blocs de 8×8 pixels deviennent apparents, comme des « artefacts » autour des contours. Le format JPEG convient donc aux photos, mais mal aux images qui comportent du texte ou des traits nets, pour lesquelles on préfère PNG.

6.8.3 Les images vectorielles

Définition 6.8 **Image vectorielle**

Une **image vectorielle** est décrite par une liste d'objets géométriques (points, segments, cercles, courbes, polygones, textes), accompagnés de leurs caractéristiques : position, dimensions, couleur et épaisseur du trait, couleur de remplissage.

Une image vectorielle ne contient aucun pixel : c'est une *recette* que le logiciel exécute au moment de l'affichage, en calculant quels pixels de l'écran colorer. Ce calcul s'appelle le *rendu*, ou la *matricialisation*. Comme il est refait à chaque affichage, à la taille demandée, l'image peut être agrandie autant qu'on le veut sans perte de qualité : un cercle reste un cercle parfait.

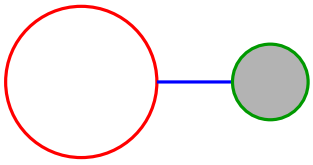


FIGURE 6.10 : Une image vectorielle très simple : deux cercles et un segment.

L'image de la figure 6.10 est décrite, en TikZ, le langage graphique utilisé pour les figures de ce livre, par les trois lignes suivantes :

```
\draw[very thick, red] (0,0) circle (1cm);
\draw[very thick, blue] (1,0) -- (2,0);
\draw[very thick, green!60!black, fill=black!30] (2.5,0) circle (0.5cm);
```

Sur le Web, le format vectoriel standard est SVG (*Scalable Vector Graphics*). C'est lui aussi un fichier texte, que l'on peut écrire à la main et ouvrir dans n'importe quel navigateur :

Programme 6.1 : La même image au format SVG. L'axe des ordonnées est orienté vers le bas, et les unités sont des pixels.

```
<svg xmlns="http://www.w3.org/2000/svg" width="300" height="200">
  <circle cx="100" cy="100" r="50" stroke="red"
          stroke-width="4" fill="none"/>
  <line x1="150" y1="100" x2="200" y2="100"
        stroke="blue" stroke-width="4"/>
  <circle cx="225" cy="100" r="25" stroke="#008000"
          stroke-width="4" fill="#B3B3B3"/>
</svg>
```

On remarque au passage les codes de couleur hexadécimaux de la section 6.3.4. Le fichier SVG occupe quelques centaines d'octets, quelle que soit la taille à laquelle on affiche l'image. Les formats PDF et EPS sont aussi vectoriels, même s'ils peuvent contenir des images matricielles : c'est le cas de la figure 6.8 dans la version PDF de ce livre.

Remarque 6.4 **Les lettres que vous lisez sont des images vectorielles**

Les polices de caractères modernes (TrueType, OpenType) décrivent chaque lettre par son contour, formé de segments et de *courbes de Bézier*, des courbes paramétrées polynomiales de degré 2 ou 3 définies par quelques points de contrôle. C'est ce qui permet d'afficher le même texte, avec la même netteté, en petits caractères sur un téléphone et en lettres géantes sur une affiche. Le logiciel de dessin vectoriel libre Inkscape permet de manipuler ces points de contrôle et d'en observer l'effet.

Les deux familles se prêtent différemment aux transformations géométriques du chapitre 1. Pour appliquer une transformation affine à une image vectorielle, il suffit de l'appliquer aux points qui définissent ses objets : le résultat est exact. Pour une image matricielle, il faut *rééchantillonner* : chaque pixel de l'image d'arrivée provient d'un point de l'image de départ qui ne tombe généralement pas au centre d'un pixel, et dont il faut estimer la couleur à partir de ses voisins. Faire tourner une photo de quelques degrés, puis la remettre droite, la rend légèrement floue.

Table 6.7 : Les images matricielles et vectorielles en résumé.

	Image matricielle	Image vectorielle
Description	un tableau de pixels	une liste d'objets géométriques
Agrandissement	les pixels deviennent visibles	aucune perte de qualité
Taille du fichier	dépend de la définition	dépend du nombre d'objets
Usages	photos, images numérisées	logos, plans, schémas, polices
Formats	BMP, PNG, JPEG, GIF	SVG, PDF, EPS
Logiciels	GIMP, Photoshop	Inkscape, Illustrator, GeoGebra, TikZ

Le choix dépend donc de la nature de l'image. Une photo n'a pas de description géométrique simple, et l'on ne peut que l'échantillonner : c'est le domaine des images matricielles. Un logo, un plan ou une figure géométrique sont, par nature, faits d'objets géométriques : les décrire par des pixels serait à la fois plus lourd et moins précis. Pour les enseignants de mathématiques, il y a là une conséquence pratique : les figures produites avec GeoGebra et destinées à un document imprimé gagnent à être exportées en SVG ou en PDF plutôt qu'en PNG.

Exercices

Exercice 6.1. Chacune des chaînes de 8 bits suivantes est l'écriture en base deux d'un entier compris entre 0 et 255. Trouver cet entier.

a) `00010000`

b) `10101010`

c) `00001111`

d) `11111011`

Exercice 6.2. Écrire chacun des entiers suivants sous la forme d'une chaîne de 8 bits.

a) 0
b) 21
c) 144
d) 251

Exercice 6.3. Écrire en base seize les nombres suivants, donnés en base dix.

a) 10
b) 100
c) 243
d) 9999

Exercice 6.4. Écrire en base dix les nombres suivants, donnés en base seize.

a) `123`
b) `1AC`
c) `ABBA`
d) `9C9A`

Exercice 6.5. On utilise la proposition 6.2.

a) Écrire $(1011\,0110\,1110)_2$ en base huit et en base seize, sans passer par la base dix.
b) Écrire $(7\mathrm{EA})_{16}$ en base deux, puis en base huit.
c) Pourquoi ne peut-on pas convertir aussi simplement de la base deux à la base dix ? Peut-on convertir par regroupement de la base trois à la base neuf ?

Exercice 6.6. Combien d'octets faut-il pour stocker un nombre écrit avec 16 chiffres hexadécimaux ? Combien de valeurs différentes ces chiffres permettent-ils de représenter ?

Exercice 6.7. On utilise la formule $\lfloor \log_b n \rfloor + 1$ donnant le nombre de chiffres de n en base b.

a) Combien de chiffres faut-il pour écrire 10^{100} en base deux ? en base seize ?
b) Montrer qu'un entier écrit en base dix avec k chiffres s'écrit en base deux avec au plus $\lfloor k \log_2 10 \rfloor + 1$ bits.

Exercice 6.8. Cet exercice utilise les congruences de la section E.6.

a) Montrer qu'un entier est divisible par 8 si et seulement si les trois derniers chiffres de son écriture binaire sont nuls.
b) Montrer qu'un entier est divisible par 15 si et seulement si la somme de ses chiffres hexadécimaux est divisible par 15. Quelle règle de la base dix cette propriété généralise-t-elle ?
c) Le nombre $(\mathrm{ABBA})_{16}$ est-il divisible par 3 ? par 5 ?

Exercice 6.9. On reprend la remarque 6.1. On suppose que l'ordinateur du Patriot conserve 0,1 en ne gardant que les 23 premiers bits après la virgule de son développement binaire, et en tronquant le reste.

a) À l'aide du développement $0{,}1 = (0{,}0\,0011\,0011\,\ldots)_2$, montrer que l'erreur commise est $0{,}1 \cdot 2^{-20} \approx 9{,}5 \times 10^{-8}$.
b) Quelle est l'erreur accumulée sur l'horloge après 100 heures de fonctionnement ?
c) Un Scud se déplace à environ 1676 mètres par seconde. De combien de mètres la position prédite est-elle erronée ?

Exercice 6.10. On utilise les codes de couleur de la section 6.3.4.

a) Quelles couleurs sont décrites par les codes `#00FF00`, `#000080`, `#FFC0CB` et `#C0C0C0` ?

b) Donner le code hexadécimal de la couleur RVB $(255, 165, 0)$.

c) Donner le code de la couleur complémentaire de `#1C3F6E`.

d) Combien de couleurs peut-on écrire avec la notation abrégée à trois chiffres, comme `#F80` ?

e) Une « palette » de 256 couleurs, utilisée par le format GIF, associe à chaque pixel le *numéro* d'une couleur plutôt que la couleur elle-même. Combien d'octets économise-t-on ainsi pour une image de 640×480 pixels, en tenant compte de la place occupée par la palette ?

Exercice 6.11. On utilise les permissions Unix décrites à la section 6.3.

a) Que signifie la commande `chmod 640 notes.txt` ?

b) Quelle commande donne au propriétaire tous les droits, et aucun droit à quiconque d'autre ?

c) Combien de combinaisons de permissions différentes existe-t-il ?

Exercice 6.12. On utilise le gabarit de la définition 6.5.

a) Encoder en UTF-8 les caractères « ç » (`U+00E7`), « « » (`U+00AB`) et « œ » (`U+0153`). Donner le résultat en binaire et en hexadécimal.

b) Décoder la suite d'octets `43 61 66 C3 A9`.

c) Combien d'octets occupe la phrase « Ça va ? » en UTF-8 ? en Windows-1252 ?

d) Un texte UTF-8 contenant le mot « garçon » est lu comme du Windows-1252. Qu'affiche-t-on ?

e) Combien de bits utiles transporte un caractère de 4 octets ? Expliquer pourquoi l'espace des points de code d'Unicode, qui va jusqu'à `U+10FFFF`, tient dans ce gabarit.

Exercice 6.13. On distingue soigneusement les préfixes décimaux et binaires, ainsi que les bits et les octets.

a) Un disque dur est vendu comme un disque de 2 To. Quelle capacité Windows affichera-t-il ?

b) Combien de temps faut-il au minimum pour télécharger un film de 4 Go avec une connexion de 50 Mbit/s ?

c) Une clé USB de « 64 Go » contient-elle assez d'espace pour 10 000 photos de 4,5 Mo chacune ?

d) Pour quelle puissance k l'écart entre 2^{10k} et 10^{3k} dépasse-t-il 20 % ?

Exercice 6.14. Pour chaque type d'image décrit ci-dessous, calculer la taille des données, en octets, d'une image de 1024×1280 pixels.

a) Une image en noir et blanc.

b) Une image en 256 tons de gris.

c) Une image en couleurs RVB, où chaque composante peut prendre 256 valeurs.

d) L'image de la question c), après une compression de rapport $4 : 1$.

Exercice 6.15. On compresse les lignes d'une image en noir et blanc par codage par plages : chaque ligne est décrite par la longueur de ses plages successives, en commençant toujours par une plage de pixels blancs (éventuellement de longueur nulle).

a) Coder la ligne `000011111111000000000011`.
b) Décoder la ligne « 0, 3, 5, 2 ».
c) Donner un exemple de ligne de 16 pixels pour laquelle le codage par plages est inefficace.

Exercice 6.16. Répondre en quelques phrases.

a) Quelle est la différence entre une image vectorielle et une image matricielle ? Donner deux exemples d'usage de chacune.
b) Pourquoi une image au format BMP est-elle plus volumineuse, à définition égale, qu'une image au format JPEG ?
c) Une figure géométrique conçue dans GeoGebra doit être imprimée sur une affiche de 1 mètre de large. Faut-il l'exporter en PNG ou en SVG ? Pourquoi ?

Exercice 6.17. On programme les deux conversions de la section 6.1.

a) Écrire une fonction `ecrire(n, b)` qui renvoie la liste des chiffres de l'entier $n \geq 0$ en base b, à l'aide de l'algorithme 6.1.
b) Écrire la fonction réciproque `lire(chiffres, b)`, à l'aide du schéma de Horner.
c) Vérifier les deux fonctions sur les écritures de 2026 en bases 2, 8 et 16, et les comparer avec `bin`, `oct` et `hex`.

Exercice 6.18. Écrire une fonction `utf8(p)` qui renvoie la liste des octets de l'encodage UTF-8 du point de code `p`, en appliquant le gabarit de la définition 6.5. La comparer avec la méthode `encode` sur les caractères `"A"`, `"é"`, `"€"` et `chr(0x1F600)`.

Exercice 6.19. Le format PPM « texte » (en-tête `P3`) décrit une image en couleurs : la première ligne est `P3`, la deuxième donne la largeur et la hauteur, la troisième la valeur maximale d'une composante (ici 255), puis viennent les composantes R, V et B de chaque pixel, ligne par ligne.

a) Écrire un programme qui produit un fichier `degrade.ppm` de 256×256 pixels, où le pixel de la colonne x et de la ligne y a pour couleur $(x, y, 128)$. Ouvrir le fichier avec GIMP.
b) Quelle est la taille du fichier produit ? La comparer avec la taille des données brutes, $256 \times 256 \times 3$ octets, et expliquer l'écart.

Exercice 6.20. Ce chapitre répond à une question que les élèves posent souvent : comment l'information est-elle stockée dans un ordinateur ?

a) Relever trois notions du programme du secondaire [Min16] qui interviennent dans ce chapitre, et préciser le rôle de chacune.
b) Un élève affirme : « Un ordinateur, c'est juste des 0 et des 1, parce que le courant passe ou ne passe pas. » Rédiger une réponse de quelques lignes, adaptée à un élève de quatrième secondaire, qui nuance cette affirmation sans la rejeter.
c) Concevoir les grandes lignes d'une activité d'une période qui amène les élèves à découvrir la notion de base de numération à partir des codes de couleur du Web.

7 La cryptographie

Chaque fois que nous payons en ligne, que nous consultons nos courriels ou que nous envoyons un message sur notre téléphone, un petit cadenas apparaît quelque part, et des calculs se font en silence. Ces calculs ont un but très ancien : permettre à deux personnes de communiquer sans qu'une troisième, qui écoute, puisse comprendre ce qu'elles se disent. Pendant des siècles, ce besoin a été celui des rois, des généraux et des espions. Il est aujourd'hui celui de tout le monde.

L'art de cacher le sens d'un message s'appelle la *cryptographie*, et son histoire est une longue course entre ceux qui inventent des chiffrements et ceux qui les cassent. Ce chapitre suit cette course. Nous commencerons par les chiffrements les plus simples, ceux de César et le chiffre affine, qui ne demandent que de l'arithmétique modulaire et des fonctions du premier degré. Nous verrons comment l'*analyse des fréquences*, inventée au IXe siècle par le savant Al-Kindi, permet de les casser, et comment Blaise de Vigenère, au XVIe siècle, a tenté d'y échapper. Le chiffre de Hill nous montrera que les matrices, vues au chapitre 1, ont aussi leur place en cryptographie. Nous ferons ensuite un détour par la machine Enigma, dont le décryptage pendant la Seconde Guerre mondiale a donné naissance aux premiers ordinateurs.

La seconde partie du chapitre raconte une révolution. Jusqu'en 1976, tout le monde pensait que deux personnes devaient partager un secret, la clé, *avant* de pouvoir communiquer en secret. La *cryptographie à clé publique* a montré que c'était faux. Nous étudierons l'échange de clés de Diffie et Hellman, puis le chiffrement RSA, qui repose directement sur le théorème d'Euler. Les résultats d'arithmétique utilisés (divisibilité, PGCD, algorithme d'Euclide, inverses modulaires, théorèmes de Fermat et d'Euler) sont rassemblés, avec leurs démonstrations, dans l'annexe E ; nous y renverrons au besoin plutôt que de les répéter. Les programmes sont écrits en Python (annexe A).

7.1 Le vocabulaire de la cryptologie

7.1.1 Chiffrer, déchiffrer, décrypter

La tradition veut que l'on appelle *Alice* la personne qui envoie un message, *Bob* celle qui le reçoit, et *Ève* l'indiscrète qui écoute (de l'anglais *eavesdropper*). Alice veut transmettre un message, appelé le **texte clair**. Elle le transforme en un **texte chiffré** (ou **cryptogramme**), incompréhensible pour qui ne connaît pas le secret, et l'envoie par un canal où Ève peut le lire. Cette transformation est le **chiffrement**. Bob, qui connaît le secret, effectue l'opération inverse, le **déchiffrement**, et retrouve le texte clair. Le secret partagé par Alice et Bob est la **clé**.

Ève, elle, ne connaît pas la clé. Si elle parvient tout de même à retrouver le texte clair, on dit qu'elle a **décrypté** le message, ou qu'elle a **cassé** le chiffrement. La **cryptographie** est l'art de concevoir des chiffrements, et la **cryptanalyse**, celui de les casser ; les deux

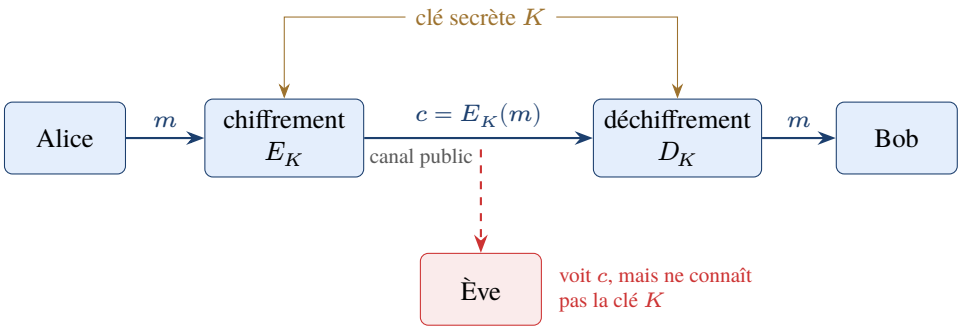


FIGURE 7.1 : Le schéma d'une communication chiffrée. Alice et Bob partagent une clé secrète K. Ève peut intercepter le texte chiffré c, mais, sans la clé, elle ne devrait pas pouvoir retrouver le texte clair m.

forment la **cryptologie**. En français, on évite le mot « crypter » et on parle de chiffrer ; « décrypter » n'est pas un synonyme de « déchiffrer », puisqu'on décrypte précisément *sans* la clé.

On ne confondra pas non plus chiffrement et *code*. Un code remplace des mots ou des expressions entières par d'autres, à l'aide d'un dictionnaire : « l'aigle a atterri » pour « l'opération a réussi ». Un chiffrement agit sur les lettres (ou sur les bits) du message, sans tenir compte de son sens. C'est ce second procédé, beaucoup plus mathématique, qui nous intéressera.

Définition 7.1 **Cryptosystème**

Un **cryptosystème** est formé d'un ensemble $\mathcal{M}$ de textes clairs, d'un ensemble $\mathcal{C}$ de textes chiffrés et d'un ensemble $\mathcal{K}$ de clés, ainsi que, pour chaque clé $K \in \mathcal{K}$, d'une fonction de chiffrement $E_K : \mathcal{M} \to \mathcal{C}$ et d'une fonction de déchiffrement $D_K : \mathcal{C} \to \mathcal{M}$ telles que

$$D_K(E_K(m)) = m \qquad \text{pour tout } m \in \mathcal{M}.$$

La condition $D_K \circ E_K = \mathrm{id}$ exprime simplement que Bob doit pouvoir retrouver le message. Elle entraîne que E_K est *injective* : deux messages différents ne peuvent pas avoir le même chiffré. Lorsque $\mathcal{M} = \mathcal{C}$ est un ensemble fini, comme ce sera souvent le cas, E_K est donc une bijection, et D_K est sa réciproque.

7.1.2 Le principe de Kerckhoffs

Que doit-on garder secret : la méthode de chiffrement, ou seulement la clé ? En 1883, le linguiste et cryptographe Auguste Kerckhoffs a donné une réponse qui fait encore autorité [Ker83] : un cryptosystème doit rester sûr même si l'ennemi en connaît tout, sauf la clé. Une méthode finit toujours par être connue (un appareil est capturé, un employé change de camp), alors qu'une clé se change facilement. Les méthodes de chiffrement utilisées aujourd'hui sur Internet sont d'ailleurs publiques, étudiées et attaquées par des chercheurs du monde entier ; c'est justement parce qu'elles ont résisté à ces attaques qu'on leur fait confiance.

Dans tout ce chapitre, nous supposerons donc qu'Ève connaît la méthode utilisée et cherche la clé. Une première conséquence est immédiate : *il doit y avoir beaucoup de clés*. Sinon, Ève les essaie toutes, ce que l'on appelle une **attaque exhaustive** (ou attaque par force brute). Ce n'est toutefois pas suffisant, comme nous le verrons à la section 7.4.

7.1.3 Des lettres aux nombres

Pour faire des mathématiques avec des lettres, on les remplace par des nombres. Nous utiliserons la correspondance du tableau 7.1, qui numérote les lettres à partir de 0. Nous écrirons les textes clairs en minuscules et les textes chiffrés en majuscules. Nous ne garderons que les 26 lettres, sans accents, espaces ni ponctuation : « Rendez-vous à midi ! » devient `rendezvousamidi`. Les espaces trahiraient la longueur des mots, ce qui faciliterait grandement le travail d'Ève. Les textes chiffrés sont souvent présentés en groupes de cinq lettres, pour faciliter leur lecture.

TABLE 7.1 : La correspondance entre les lettres et les entiers de 0 à 25.

a	b	c	d	e	f	g	h	i	j	k	l	m
0	1	2	3	4	5	6	7	8	9	10	11	12
n	o	p	q	r	s	t	u	v	w	x	y	z
13	14	15	16	17	18	19	20	21	22	23	24	25

Lorsqu'on dépasse z, on revient à a : après 25 vient 0. C'est exactement l'arithmétique modulo 26 (section E.6). Dans ce chapitre, sauf mention contraire, $\mathbb{Z}_{26}$ désigne l'ensemble $\{0, 1, \ldots, 25\}$ muni de l'addition et de la multiplication modulo 26, et l'on écrira, par exemple, $20 + 9 = 3$ dans $\mathbb{Z}_{26}$, puisque $29 \equiv 3 \pmod{26}$.

Pour la classe 7.1 **L'arithmétique de l'horloge**

> Les élèves pratiquent l'arithmétique modulaire sans le savoir : quelle heure sera-t-il dans 50 heures ? Quel jour de la semaine tombera le 1 er janvier prochain ? Le passage de z à a dans le chiffre de César est de même nature, et une roue de chiffrement comme celle de la figure 7.2 rend la chose tangible : on la construit en classe avec deux disques de carton de tailles différentes, superposés et percés en leur centre. Une attache parisienne (une petite attache de laiton à tête ronde et à deux pattes, que l'on passe dans le trou puis que l'on écarte au dos) les retient ensemble tout en laissant le petit disque tourner sur le grand ; un bout de ficelle noué ou une punaise plantée dans une gomme font aussi l'affaire. La notion de reste de la division euclidienne, vue au primaire, suffit pour travailler ; le symbole $\equiv$ peut venir plus tard, lorsque les élèves en ressentent le besoin.

7.2 Les chiffrements par décalage

7.2.1 Le chiffre de César

L'historien romain Suétone rapporte que Jules César, dans sa correspondance privée, remplaçait chaque lettre par celle qui se trouve trois rangs plus loin dans l'alphabet : a devient D, b devient E, et ainsi de suite, jusqu'à x, y et z, qui deviennent A, B et C [Sin99]. La figure 7.2 présente ce chiffrement sous la forme d'une roue.

Avec la numérotation du tableau 7.1, le chiffre de César est la fonction $x \mapsto x + 3$ dans $\mathbb{Z}_{26}$. Rien n'oblige à décaler de trois rangs.

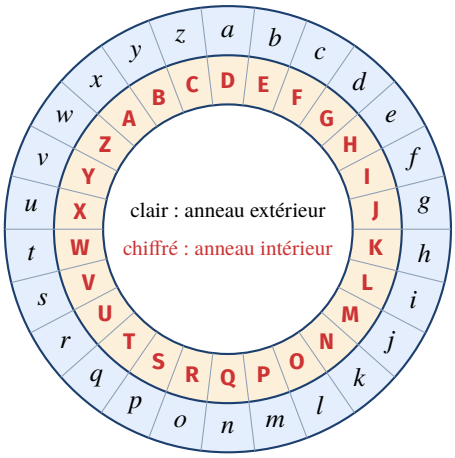


FIGURE 7.2 : Une roue de chiffrement réglée pour le chiffre de César. Chaque lettre claire, sur l'anneau extérieur, est remplacée par la lettre chiffrée placée juste en dessous. En faisant tourner l'anneau intérieur, on obtient les 26 décalages possibles.

Définition 7.2 **Chiffrement par décalage**

Soit $k \in \mathbb{Z}_{26}$. Le **chiffrement par décalage** de clé k remplace chaque lettre x du texte clair par

$$E_k(x) = x + k \bmod 26.$$

Le déchiffrement est $D_k(y) = y - k \bmod 26$. Le chiffre de César correspond à $k = 3$.

Exemple 7.1 **Alea jacta est**

Chiffrons `aleajactaest` avec la clé $k = 3$. La lettre a (0) devient D (3), l (11) devient O (14), e (4) devient H (7), et ainsi de suite ; la lettre t (19) devient W (22). On obtient `DOHDMDFWDHVW`. Pour déchiffrer, Bob recule de trois rangs.

On peut toujours ramener le déchiffrement à un chiffrement : reculer de k rangs revient à avancer de $26-k$ rangs. Un cas particulier amusant est $k = 13$: comme $13+13 = 26 \equiv 0$, chiffrer deux fois redonne le texte clair. Ce chiffrement, appelé ROT13, sert encore sur certains forums pour masquer la réponse d'une devinette ou la fin d'un film.

7.2.2 Une attaque exhaustive

Le chiffrement par décalage n'a que 26 clés, dont une, $k = 0$, ne chiffre rien du tout. Ève peut donc les essayer toutes : parmi les 26 textes obtenus, un seul, en général, aura un sens. Sur un message assez long, il n'y a aucune hésitation possible.

Exemple 7.2 **Essayer toutes les clés**

Ève intercepte `LEELBFPKOPXLTY`. Elle recule le texte de k rangs, pour $k = 0, 1, \ldots, 25$:

0 : `leelbfpkopxlty`	7 : `exxeuyidhiqemr`	14 : `xqqxnrbwabjxfk`	21 : `qjjqgkuptucqyd`
1 : `kddkaeojnowksx`	8 : `dwwdtxhcghpdlq`	15 : `wppwmqavzaiwej`	22 : `piipfjtostbpxc`
2 : `jccjzdnimnvjrw`	9 : `cvvcswgbfgockp`	16 : `voovlpzuyzhvdi`	23 : `ohhoeisnrsaowb`
3 : `ibbiycmhlmuiqv`	10 : `buubrvfaefnbjo`	17 : `unnukoytxyguch`	24 : `nggndhrmqrznva`
4 : `haahxblgklthpu`	11 : `attaquezdemain`	18 : `tmmtjnxswxftbg`	25 : `mffmcgqlpqymuz`
5 : `gzzgwakfjksgot`	12 : `zsszptdycdlzhm`	19 : `sllsimwrvwesaf`	
6 : `fyyfvzjeijrfns`	13 : `yrryoscxbckygl`	20 : `rkkrhlvquvdrze`	

Une seule ligne a un sens : avec $k = 11$, on lit « attaquez demain ». Ève a trouvé la clé et le message en quelques secondes.

Pour Ève, la seule difficulté est de reconnaître le bon texte parmi les autres. Un ordinateur le fait aussi : il suffit de retenir le décalage qui produit le plus de lettres fréquentes en français, une idée que nous développerons à la section 7.4.

7.3 Le chiffre affine

7.3.1 Chiffrer avec une fonction affine

Pour augmenter le nombre de clés, on peut remplacer l'addition par une fonction affine, la fonction du premier degré bien connue des élèves du secondaire, mais calculée modulo 26.

Définition 7.3 **Chiffre affine**

Soit $\alpha, \beta \in \mathbb{Z}_{26}$. Le **chiffre affine** de clé (α, β) remplace chaque lettre x du texte clair par

$$E_{\alpha,\beta}(x) = \alpha x + \beta \bmod 26.$$

Le chiffrement par décalage est le cas particulier $\alpha = 1$.

Exemple 7.3 **Chiffrer avec la clé $(7, 3)$**

Chiffrons le mot `tableau` avec $\alpha = 7$ et $\beta = 3$. La lettre t vaut 19, et $7 \times 19 + 3 = 136 = 5 \times 26 + 6$: elle devient 6, soit G. De même,

clair	t	a	b	l	e	a	u
x	19	0	1	11	4	0	20
$7x + 3$	136	3	10	80	31	3	143
$7x + 3 \bmod 26$	6	3	10	2	5	3	13
chiffré	G	D	K	C	F	D	N

Le texte chiffré est `GDKCFDN`. Les deux a donnent la même lettre D : comme tous les chiffrements de cette section, le chiffre affine est une *substitution monoalphabétique*, où chaque lettre est toujours remplacée par la même.

7.3.2 Quelles clés sont admissibles ?

Toutes les valeurs de α ne conviennent pas. Avec $\alpha = 2$ et $\beta = 0$, la lettre a (0) et la lettre n (13) deviennent toutes les deux A, puisque $2 \times 13 = 26 \equiv 0$; plus généralement, x et $x + 13$ ont toujours le même chiffré, et le texte chiffré `ACE` peut provenir de `abc`, de `nbc`, de `apc`… Bob ne peut pas déchiffrer. La proposition suivante dit exactement quand le chiffrement est inversible.

Proposition 7.1 **Clés admissibles du chiffre affine**

La fonction $E_{\alpha,\beta}(x) = \alpha x + \beta$ est une bijection de $\mathbb{Z}_{26}$ dans lui-même si et seulement si $\text{pgcd}(\alpha, 26) = 1$. Dans ce cas, sa réciproque est

$$D(y) = \alpha'(y - \beta) = \alpha' y - \alpha'\beta \bmod 26,$$

où α' est l'inverse de α modulo 26. Le déchiffrement est donc lui-même un chiffre affine, de clé $(\alpha', -\alpha'\beta)$.

Démonstration. Si $\text{pgcd}(\alpha, 26) = 1$, l'entier α possède un inverse α' modulo 26 (proposition E.19). Alors, pour tout x,

$$D(E_{\alpha,\beta}(x)) = \alpha'(\alpha x + \beta - \beta) = \alpha'\alpha\, x \equiv x \pmod{26},$$

et de même $E_{\alpha,\beta}(D(y)) \equiv y$. La fonction $E_{\alpha,\beta}$ est donc bijective, de réciproque D.

Si au contraire $d = \text{pgcd}(\alpha, 26) > 1$, posons $x_1 = 26/d$, un entier compris entre 1 et 25. Comme $\alpha = d\alpha_1$ pour un certain entier α_1, on a $\alpha x_1 = \alpha_1 \cdot 26 \equiv 0 \pmod{26}$, d'où $E_{\alpha,\beta}(x_1) = \beta = E_{\alpha,\beta}(0)$. Deux lettres différentes ont le même chiffré, et $E_{\alpha,\beta}$ n'est pas injective. □

Comme $26 = 2 \times 13$, la condition $\text{pgcd}(\alpha, 26) = 1$ signifie que α est impair et différent de 13. Il y a donc $\varphi(26) = 12$ valeurs admissibles (section E.8) :

$$\alpha \in \{1, 3, 5, 7, 9, 11, 15, 17, 19, 21, 23, 25\},$$

et 26 valeurs de β, soit $12 \times 26 = 312$ clés, dont une, $(1, 0)$, qui ne chiffre rien. C'est mieux que 26, mais un ordinateur les essaie toutes en une fraction de seconde.

Exemple 7.4 **Déchiffrer**

Bob reçoit `GDKCFDN`, chiffré avec la clé $(7, 3)$. L'inverse de 7 modulo 26 est 15, puisque $7 \times 15 = 105 = 4 \times 26 + 1$ (exemple E.7). La clé de déchiffrement est donc $(15, -15 \times 3) = (15, -45) \equiv (15, 7)$, c'est-à-dire $D(y) = 15y + 7$. Pour la première lettre, $G = 6$ donne $15 \times 6 + 7 = 97 = 3 \times 26 + 19$, soit t. On retrouve ainsi `tableau`.

Le programme 7.1 réalise ces calculs. La fonction `pow(a, -1, 26)` de Python (version 3.8 et suivantes) calcule directement l'inverse modulaire, à l'aide de l'algorithme d'Euclide étendu.

Programme 7.1 : Le chiffre affine (et le chiffrement par décalage, lorsque `alpha = 1`).

```
ALPHABET = "abcdefghijklmnopqrstuvwxyz"

def vers_nombres(texte):
```

```
    """Garde les lettres de a à z et les remplace par 0, 1, ..., 25."""
    return [ALPHABET.index(c) for c in texte.lower() if c in ALPHABET]

def vers_lettres(nombres):
    return "".join(ALPHABET[x % 26] for x in nombres).upper()

def affine(texte, alpha, beta):
    """Chiffre affine x -> alpha x + beta (mod 26)."""
    return vers_lettres([alpha * x + beta for x in vers_nombres(texte)])

def cle_dechiffrement(alpha, beta):
    a = pow(alpha, -1, 26)          # inverse de alpha modulo 26
    return a, (-a * beta) % 26

print(affine("Alea jacta est", 1, 3))        # DOHDMDFWDHVW
print(affine("tableau", 7, 3))               # GDKCFDN
print(cle_dechiffrement(7, 3))               # (15, 7)
print(affine("GDKCFDN", 15, 7))              # TABLEAU
for k in range(26):                          # attaque exhaustive
    print(k, affine("LEELBFPKOPXLTY", 1, -k))
```

Les accents posent un petit problème : `"é"` n'est pas dans `ALPHABET` et serait simplement supprimé. On peut d'abord retirer les accents à l'aide du module `unicodedata`, qui décompose « é » en « e » suivi d'un accent (section 6.4) : `"".join(c for c in unicodedata.normalize("NFD", texte) if c.isascii())`.

7.3.3 Une attaque à texte clair connu

Ève n'a pas toujours besoin d'essayer toutes les clés. Il arrive qu'elle connaisse, ou devine, une partie du texte clair : une formule de politesse, un en-tête, une date. Elle mène alors une **attaque à texte clair connu**. Pour le chiffre affine, deux lettres suffisent souvent, car la clé est solution d'un système de deux équations.

Exemple 7.5 **Deux lettres suffisent (presque)**

Ève intercepte `PCRZCQJ` et sait que le message est une salutation. Elle devine `bonjour`. Les lettres $b = 1$ et $u = 20$ sont devenues $P = 15$ et $Q = 16$, d'où le système

$$\alpha + \beta \equiv 15, \qquad 20\alpha + \beta \equiv 16 \pmod{26}.$$

En soustrayant, $19\alpha \equiv 1 \pmod{26}$: α est l'inverse de 19, soit 11 (en effet, $19 \times 11 = 209 = 8 \times 26 + 1$). Alors $\beta \equiv 15 - 11 = 4$. Ève vérifie avec les autres lettres : $o = 14$ donne $11 \times 14 + 4 = 158 \equiv 2$, soit C. La clé est $(11, 4)$.
Si Ève avait choisi les lettres b et o, elle aurait obtenu $13\alpha \equiv 15 - 2 = 13 \pmod{26}$. Comme 13 n'est pas inversible modulo 26, cette congruence a 13 solutions (tous les α impairs) : elle ne détermine pas la clé. En général, la différence de deux équations donne une congruence $(x_1 - x_2)\,\alpha \equiv y_1 - y_2$, qui se résout par la proposition E.20 ; lorsque $x_1 - x_2$ est pair, on obtient deux valeurs de α modulo 26, dont une seule est impaire.

7.3.4 Activité : retrouver la clé

L'attaque à texte clair connu se prête bien à une activité en classe. L'enseignante ou l'enseignant chiffre la liste des prénoms des élèves avec une clé affine secrète, et distribue

la liste chiffrée, dans le désordre. Les élèves connaissent les prénoms de leurs camarades : ils disposent donc d'un texte clair connu, mais ne savent pas quel prénom correspond à quel chiffré. Ils doivent d'abord apparier les deux listes, en s'appuyant sur la longueur des prénoms, les lettres doublées (`camille`, `isabelle`) ou les lettres répétées (`anna`, `emma`), puis résoudre un système de congruences pour trouver α et β.

La version web de ces notes contient une page interactive pour cette activité : on y entre une liste de prénoms, une clé admissible est tirée au hasard, et la liste chiffrée s'affiche dans le désordre. Un bouton permet de révéler la clé une fois l'enquête terminée. L'exercice 7.5 en propose une version sur papier.

Pour la classe 7.2 **Une fonction et sa réciproque**

> Le chiffre affine réinvestit plusieurs notions du secondaire. Chiffrer, c'est appliquer une fonction ; déchiffrer, c'est appliquer sa réciproque, et la proposition 7.1 rappelle qu'une fonction n'a de réciproque que si elle est bijective. Trouver la clé à partir de deux lettres, c'est résoudre un système de deux équations à deux inconnues, comme on le fait pour trouver la règle d'une fonction affine passant par deux points. La différence, et elle est riche de discussions, tient à l'arithmétique modulo 26 : on ne peut pas toujours « diviser », et une équation du premier degré peut avoir zéro, une ou plusieurs solutions. L'activité des prénoms ajoute une étape de raisonnement (apparier les deux listes) qui motive fortement les élèves : la clé leur est cachée, mais la réponse se vérifie d'un clic.

7.4 La substitution et l'analyse des fréquences

7.4.1 Un nombre astronomique de clés

Le chiffre affine n'est qu'un cas particulier. Une **substitution monoalphabétique** remplace chaque lettre par une autre selon une permutation quelconque de l'alphabet, c'est-à-dire une bijection de $\{a, b, \dots, z\}$ dans lui-même, que l'on peut écrire sous la forme d'un tableau à deux lignes :

$$\begin{pmatrix} a & b & c & d & e & f & g & h & i & j & k & l & m & n & o & p & q & r & s & t & u & v & w & x & y & z \\ Q & W & E & R & T & Y & U & I & O & P & A & S & D & F & G & H & J & K & L & Z & X & C & V & B & N & M \end{pmatrix}.$$

La clé est ici facile à retenir : c'est l'ordre des touches d'un clavier QWERTY. Combien y a-t-il de clés ? L'image de a peut être choisie de 26 façons, celle de b de 25 façons, et ainsi de suite ; par le principe multiplicatif, il y a

$$26! = 26 \times 25 \times \cdots \times 2 \times 1 \approx 4{,}03 \times 10^{26}$$

permutations. Un ordinateur qui en essaierait un milliard par seconde mettrait plus de dix milliards d'années à toutes les essayer, soit nettement plus que l'âge de l'Univers. L'attaque exhaustive est impossible. Pourtant, ce chiffrement se casse à la main, en une heure ou deux, comme le savent les amateurs de cryptogrammes des journaux.

7.4.2 L'analyse des fréquences

Le défaut de toute substitution monoalphabétique est qu'elle respecte les répétitions : si le e est la lettre la plus fréquente du texte clair, son remplaçant sera la lettre la plus fréquente du texte chiffré. Or, dans une langue donnée, les lettres n'apparaissent pas toutes

aussi souvent. La figure 7.3 montre les fréquences mesurées sur le texte des six premiers chapitres de ce livre, après avoir retiré les accents, les formules et le code, soit environ 290 000 lettres.

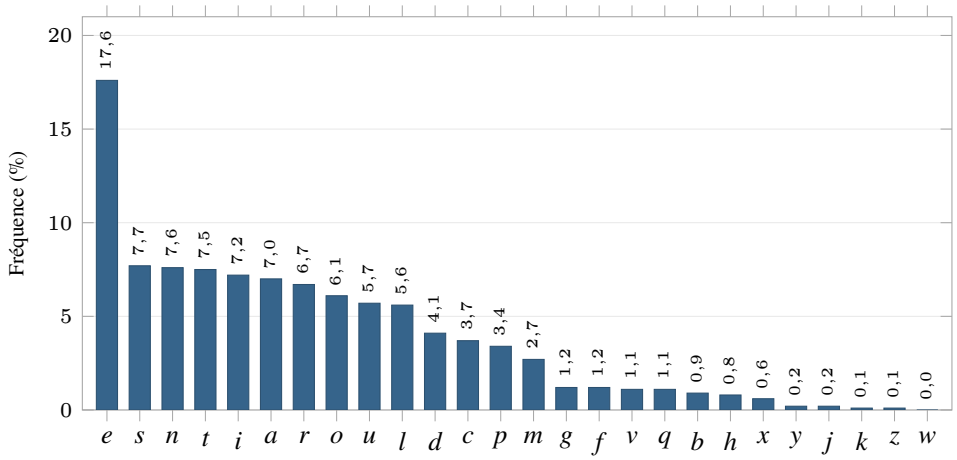


FIGURE 7.3 : Les fréquences des lettres (en pour cent) dans le texte des chapitres 1 à 6 de ce livre, une fois les accents retirés. Le e (qui regroupe ici e, , , et) représente à lui seul plus d'une lettre sur six.

Le e domine nettement, avec environ 17,6 % des lettres. Viennent ensuite, dans un ordre qui varie d'un texte à l'autre, les lettres s, n, t, i, a, r, o, u et l, toutes entre 5 % et 8 %. Les lettres k, w et z sont rarissimes en français, alors que w est fréquent en anglais : les fréquences permettent même de deviner la langue d'un texte chiffré.

Note historique 7.1 Al-Kindi, le premier cryptanalyste

Abu Yusuf Ya'qub ibn Ishaq al-Kindi (vers 801–873) vécut à Bagdad, au temps où la « Maison de la sagesse » faisait traduire en arabe les ouvrages des savants grecs. Philosophe, mathématicien, astronome et musicien, il écrivit plusieurs centaines de traités. L'un d'eux, le *Manuscrit sur le déchiffrement des messages cryptographiques*, ne fut redécouvert qu'en 1987, dans les archives ottomanes d'Istanbul. On y trouve la plus ancienne description connue de l'analyse des fréquences. Pour élucider un message chiffré dont on connaît la langue, explique Al-Kindi, on se procure un autre texte dans cette langue, long d'environ une page, et l'on y compte les apparitions de chaque lettre ; on classe ensuite de la même façon les symboles du message chiffré, et l'on remplace le plus fréquent par la lettre la plus fréquente du texte de référence, et ainsi de suite [Sin99]. L'idée tient en une phrase ; il a pourtant fallu attendre plus de six siècles pour que l'Europe de la Renaissance la redécouvre et que les substitutions monoalphabétiques soient abandonnées par les diplomates.

7.4.3 Casser un chiffre affine par les fréquences

Pour une substitution quelconque, il faut deviner les 26 lettres, en s'aidant des fréquences, des lettres doublées (`ss`, `ll`, `ee`), des petits mots et du contexte. Pour un chiffre affine, c'est beaucoup plus rapide : il suffit d'identifier *deux* lettres, puisque deux équations déterminent la clé.

Exemple 7.6 **Un message chiffré avec une clé affine inconnue**

Ève intercepte le texte suivant, qu'elle sait chiffré par un chiffre affine.

```
JOGLW TDKQL OTZIY YUDJK BQDYS TULYU DYYMB YOHHO ILYPY
PITJK SODYU YDPYS IJIDO ILYUJ YULKI UJYUQ YBYLO MNYDJ
YUYUT IKBUV KMJOI YBDTL KDYQY LJYML UUYGL YDUYD JYMLU
OPVYL UOILY UGZYL GZOIY BDUOB UGYUU YOJYU TYLGY LOMRK
MLPZM IYJJY TLKDY QYBKU OGZOD UYBJI QBYBK USYUU OQYUY
DBKUG OLDYU XOBGO ILYUG ZOCMY HKIUC MYBKM UVIUI DKBUM
BUIDY UYGML IUYPY UGOJG MJUUM LPYDL YUQLO BPUBK SXLYU
YBDIY LUUYP YLKMJ YBDYB UIJYB GYPOB UBKDL YDYJY TZKBY
```

Ce texte compte 360 lettres. Les plus fréquentes sont Y (69 fois), U (48 fois), puis L (29), B (25), O et D (24 chacune). Ève fait l'hypothèse que Y (24) chiffre e (4), et que U (20) chiffre s (18) :

$$4\alpha + \beta \equiv 24, \qquad 18\alpha + \beta \equiv 20 \pmod{26}.$$

En soustrayant, $14\alpha \equiv -4 \equiv 22 \pmod{26}$. Comme $\text{pgcd}(14, 26) = 2$ divise 22, on divise tout par 2 : $7\alpha \equiv 11 \pmod{13}$. L'inverse de 7 modulo 13 est 2, d'où $\alpha \equiv 22 \equiv 9 \pmod{13}$, c'est-à-dire $\alpha = 9$ ou $\alpha = 22$ dans $\mathbb{Z}_{26}$. Seule la valeur impaire est admissible : $\alpha = 9$, puis $\beta \equiv 24 - 36 = -12 \equiv 14$.
La clé de déchiffrement est $(9', -9' \times 14)$, où $9' = 3$ est l'inverse de 9 modulo 26 ($9 \times 3 = 27$), soit $(3, -42) \equiv (3, 10)$. Le texte déchiffré commence par `lacryptographieestlongtempsrestee…` : l'hypothèse était bonne. Si elle ne l'avait pas été, Ève aurait essayé d'associer U à une autre lettre fréquente (n, t, a…), jusqu'à obtenir un texte lisible.

Cet exemple montre que le nombre de clés n'est pas tout : un bon chiffrement doit aussi *effacer les régularités* du texte clair. C'est ce que vont tenter de faire les chiffrements des deux sections suivantes.

Pour la classe 7.3 **Statistique et cryptanalyse**

L'analyse des fréquences est une belle activité de statistique pour le premier cycle du secondaire. Les élèves dressent le tableau de distribution des lettres d'un texte de leur choix, calculent des fréquences relatives et construisent un diagramme à bandes. En comparant leurs résultats, ils constatent que les fréquences varient d'un texte à l'autre, mais d'autant moins que les textes sont longs : c'est une première rencontre avec l'idée qu'une fréquence observée estime une probabilité. Ils peuvent ensuite casser un cryptogramme préparé par l'enseignant, ce qui donne un but concret à toute la démarche. Un texte court, ou écrit volontairement sans la lettre e (comme le roman *La Disparition* de Georges Perec), montre les limites de la méthode.

Outil 7.1 **Des outils en ligne**

Le site *CrypTool-Online* (`www.cryptool.org`) propose, en français, la plupart des chiffrements de ce chapitre, avec des outils d'analyse des fréquences. En Python, la classe `Counter` du module `collections` compte les lettres d'un texte en une ligne : `Counter(vers_lettres(vers_nombres(texte))).most_common()` donne la

liste des lettres, de la plus fréquente à la moins fréquente.

7.5 Le chiffre de Vigenère

7.5.1 Plusieurs décalages

Pour échapper à l'analyse des fréquences, il faut qu'une même lettre claire ne soit pas toujours chiffrée de la même façon. L'idée la plus simple consiste à utiliser plusieurs décalages, à tour de rôle. Elle a été décrite par Giovan Battista Bellaso en 1553, puis popularisée par le diplomate français Blaise de Vigenère dans son *Traicté des chiffres* (1586), et c'est le nom de ce dernier qui est resté.

Définition 7.4 **Chiffre de Vigenère**

Soit $(k_0, k_1, \ldots, k_{\ell-1})$ une suite de ℓ éléments de $\mathbb{Z}_{26}$, appelée la clé, de **longueur** ℓ. Le **chiffre de Vigenère** remplace la lettre x_i en position i du texte clair (les positions sont numérotées à partir de 0) par

$$y_i = x_i + k_{i \bmod \ell} \bmod 26.$$

La lettre en position i est donc décalée de k_0, k_1, …, selon que i est congru à 0, 1, … modulo ℓ. Le déchiffrement soustrait les mêmes décalages.

Pour retenir la clé, on la choisit souvent sous la forme d'un mot, dont les lettres donnent les décalages.

Exemple 7.7 **Rendez-vous à minuit**

Chiffrons `rendezvousaminuit` avec le mot-clé `cle`, c'est-à-dire avec les décalages $(2, 11, 4)$. On écrit le mot-clé de façon répétée sous le texte, et on additionne les colonnes modulo 26 :

clair	r	e	n	d	e	z	v	o	u	s	a	m	i	n	u	i	t
clé	c	l	e	c	l	e	c	l	e	c	l	e	c	l	e	c	l
chiffré	T	P	R	F	P	D	X	Z	Y	U	L	Q	K	Y	Y	K	E

Le texte chiffré est `TPRFPDXZYULQKYYKE`. La lettre n apparaît deux fois dans le texte clair, et elle est chiffrée une fois en R, une fois en Y. À l'inverse, Y apparaît trois fois dans le chiffré, pour deux lettres claires différentes (u et n). Les deux e, eux, sont tous deux devenus P, parce qu'ils sont séparés par 3 positions, la longueur de la clé : c'est par ce genre de répétitions que le chiffre se trahit.

Pendant trois siècles, ce chiffrement a eu la réputation d'être inviolable : on le surnommait « le chiffre indéchiffrable ». Il a pourtant été cassé, vers 1854 par Charles Babbage, qui ne publia pas sa méthode, puis en 1863 par l'officier prussien Friedrich Kasiski [Sin99]. L'attaque se fait en deux temps : trouver la longueur de la clé, puis chacun des décalages.

7.5.2 Trouver la longueur de la clé

Supposons qu'Ève connaisse la longueur ℓ de la clé. Les lettres en positions $0, \ell, 2\ell, \ldots$ ont toutes été décalées de k_0 : elles forment un texte chiffré par un simple décalage, dont les

fréquences sont celles du français, permutées circulairement. La lettre la plus fréquente de cette colonne chiffre probablement e, ce qui donne k_0. On procède de même pour les positions $1, \ell+1, 2\ell+1, \ldots$, et ainsi de suite. Un chiffre de Vigenère de longueur ℓ n'est donc que ℓ chiffres de César entrelacés.

Il reste à trouver ℓ. Une méthode simple, due à William Friedman (1922), consiste à compter les **coïncidences**. On écrit le texte chiffré sur une bande, on en fait une copie que l'on décale de d positions, et on compte le nombre de positions où les deux bandes portent la même lettre. Si d est un multiple de ℓ, les deux lettres superposées ont subi le même décalage : elles coïncident aussi souvent que deux lettres prises au hasard dans un texte français, soit environ 7,8 % du temps avec les fréquences de la figure 7.3. (Cette probabilité est la somme des carrés des fréquences : $0{,}176^2 + 0{,}077^2 + \cdots$.) Si d n'est pas un multiple de ℓ, les décalages sont différents, et les coïncidences sont plus rares, autour de $1/26 \approx 3{,}8$ %. Le nombre de coïncidences présente donc des pics aux multiples de ℓ.

Exemple 7.8 **Le texte de l'exemple 7.6, chiffré par Vigenère**

On chiffre les 360 lettres de l'exemple 7.6 par Vigenère, avec le mot-clé `secondaire`. Les nombres de coïncidences pour les décalages $d = 1, 2, \ldots, 12$ sont

$$24,\ 17,\ 15,\ 14,\ 12,\ 19,\ 16,\ 15,\ 13,\ \mathbf{43},\ 18,\ 10.$$

Le pic à $d = 10$ (43 coïncidences sur 350 comparaisons, soit 12,3 %, contre 3 à 7 % ailleurs) révèle la longueur de la clé. Il reste à analyser dix colonnes de 36 lettres chacune. Le programme 7.2 retrouve alors le mot `SECONDAIRE`.

Les colonnes sont courtes, et la lettre la plus fréquente d'une colonne n'est pas toujours le chiffré de e. Plutôt que de ne regarder que le maximum, le programme 7.2 essaie les 26 décalages de chaque colonne et retient celui qui donne le texte le « plus français », c'est-à-dire celui pour lequel la somme des fréquences françaises des lettres obtenues est la plus grande.

Programme 7.2 : Le chiffre de Vigenère et sa cryptanalyse (on utilise les fonctions du programme 7.1).

```
def vigenere(texte, cle, signe=1):
    """Chiffre (signe = 1) ou déchiffre (signe = -1) avec le mot-clé cle.
    """
    k = vers_nombres(cle)
    x = vers_nombres(texte)
    return vers_lettres([x[i] + signe * k[i % len(k)] for i in range(len
    (x))])

def coincidences(texte, d):
    """Nombre de positions où le texte et sa copie décalée de d
    coïncident."""
    x = vers_nombres(texte)
    return sum(x[i] == x[i + d] for i in range(len(x) - d))

# Fréquences du français (en %), de a à z (figure des fréquences)
FREQ = [7.0, 0.9, 3.7, 4.1, 17.6, 1.2, 1.2, 0.8, 7.2, 0.2, 0.1, 5.6, 2.7,

        7.6, 6.1, 3.4, 1.1, 6.7, 7.7, 7.5, 5.7, 1.1, 0.0, 0.6, 0.2, 0.1]

def meilleur_decalage(colonne):
    """Le décalage k qui rend la colonne la plus semblable à du français.
```

```
    """
    return max(range(26), key=lambda k: sum(FREQ[(y - k) % 26] for y in
    colonne))

def trouver_cle(chiffre, longueur):
    y = vers_nombres(chiffre)
    return vers_lettres([meilleur_decalage(y[j::longueur]) for j in
    range(longueur)])

c = vigenere("rendez-vous à minuit", "cle")
print(c, vigenere(c, "cle", -1))     # TPRFPDXZYULQKYYKE
    RENDEZVOUSAMINUIT
```

L'expression `y[j::longueur]` extrait les éléments de la liste `y` en positions j, $j + \ell$, $j + 2\ell$, …, c'est-à-dire la j-ième colonne.

7.5.3 Le masque jetable

Les attaques précédentes exploitent la périodicité de la clé. Que se passe-t-il si la clé est aussi longue que le message, choisie entièrement au hasard, et jamais réutilisée ? On obtient le **masque jetable**, proposé par Gilbert Vernam et Joseph Mauborgne vers 1917. Claude Shannon a démontré en 1949 que ce chiffrement est *parfaitement sûr* : un texte chiffré de n lettres peut provenir de n'importe quel texte clair de n lettres, chacun avec la même probabilité, puisqu'il existe exactement une clé qui transforme l'un en l'autre. Ève, même dotée d'une puissance de calcul infinie, n'apprend rien d'autre que la longueur du message.

Le prix à payer est lourd : il faut produire, transporter et garder secrète une clé aussi longue que tous les messages que l'on échangera. Le masque jetable a été utilisé pour des communications diplomatiques et militaires de très haute importance, mais il est inutilisable pour les milliards de connexions quotidiennes d'Internet. Il pose de façon aiguë le problème qui nous occupera à la section 7.8 : comment Alice et Bob peuvent-ils se mettre d'accord sur une clé ?

7.6 Le chiffre de Hill

7.6.1 Chiffrer par blocs avec une matrice

Les chiffrements précédents traitent les lettres une à une. En 1929, le mathématicien américain Lester S. Hill a proposé de chiffrer des *blocs* de lettres à l'aide d'une matrice [Hil29]. On découpe le texte clair en blocs de n lettres, que l'on voit comme des vecteurs de $\mathbb{Z}_{26}^n$, et l'on multiplie chaque vecteur par une même matrice A de taille $n \times n$, en calculant modulo 26.

Définition 7.5 **Chiffre de Hill**

Soit A une matrice $n \times n$ à coefficients entiers. Le **chiffre de Hill** de clé A découpe le texte clair en blocs $\mathbf{x} = (x_1, \dots, x_n)$ de n lettres (en complétant le dernier bloc au besoin, par exemple avec des x) et remplace chaque bloc par

$$\mathbf{y} = A\mathbf{x} \bmod 26,$$

où le produit matrice-vecteur est réduit coefficient par coefficient modulo 26.

Exemple 7.9 **Un chiffre de Hill 2×2**

Prenons

$$A = \begin{pmatrix} 3 & 2 \\ 5 & 7 \end{pmatrix}$$

et chiffrons `cryptage`. Les blocs sont `cr`, `yp`, `ta` et `ge`, soit les vecteurs $(2, 17)$, $(24, 15)$, $(19, 0)$ et $(6, 4)$. Pour le premier,

$$\begin{pmatrix} 3 & 2 \\ 5 & 7 \end{pmatrix} \begin{pmatrix} 2 \\ 17 \end{pmatrix} = \begin{pmatrix} 6 + 34 \\ 10 + 119 \end{pmatrix} = \begin{pmatrix} 40 \\ 129 \end{pmatrix} \equiv \begin{pmatrix} 14 \\ 25 \end{pmatrix} \pmod{26},$$

ce qui donne `OZ`. Les autres blocs donnent `YR`, `FR` et `AG`, et le texte chiffré est `OZYRFRAG`.

Le chiffre de Hill est le premier chiffrement *polygraphique* dont le fonctionnement soit entièrement algébrique. Il a deux qualités qui manquent aux chiffrements précédents. D'abord, il brouille les fréquences : le *a* du bloc `ta` devient R, celui du bloc `ag` d'un autre texte deviendrait autre chose, puisque l'image d'une lettre dépend de sa voisine. Ensuite, il réalise une certaine *diffusion* : modifier une seule lettre du texte clair modifie en général toutes les lettres du bloc chiffré. On retrouve ici les applications linéaires du chapitre 1, à ceci près que le « plan » est maintenant la grille finie $\mathbb{Z}_{26} \times \mathbb{Z}_{26}$.

7.6.2 Inverser une matrice modulo 26

Pour déchiffrer, Bob doit résoudre $A\mathbf{x} \equiv \mathbf{y}$, donc disposer d'une matrice B telle que $BA \equiv I \pmod{26}$. Dans $\mathbb{R}$, une matrice est inversible lorsque son déterminant n'est pas nul. Modulo 26, il faut davantage : il faut pouvoir *diviser* par le déterminant.

Théorème 7.2 **Matrice inversible modulo m**

Soit A une matrice carrée à coefficients entiers, et $m \geq 2$ un entier. Il existe une matrice B à coefficients entiers telle que $AB \equiv BA \equiv I \pmod{m}$ si et seulement si $\operatorname{pgcd}(\det A, m) = 1$. Dans ce cas, on peut prendre

$$B = \delta \operatorname{adj}(A) \bmod m,$$

où δ est l'inverse de $\det A$ modulo m et $\operatorname{adj}(A)$ est la matrice adjointe de A, c'est-à-dire la transposée de la matrice des cofacteurs.

Démonstration. Rappelons que $A \operatorname{adj}(A) = \operatorname{adj}(A)\, A = (\det A)\, I$. Cette identité classique, qui découle du développement par cofacteurs (annexe D), n'utilise que des sommes et des produits : elle est donc vraie entre matrices à coefficients entiers. Si $\operatorname{pgcd}(\det A, m) =$

1, soit δ l'inverse de $\det A$ modulo m (proposition E.19). Alors

$$A(\delta \operatorname{adj}(A)) = \delta(\det A)\, I \equiv I \pmod{m},$$

et de même dans l'autre ordre.

Réciproquement, si $AB \equiv I \pmod{m}$, alors $AB = I + mC$ pour une certaine matrice entière C. Le déterminant d'une matrice entière se calcule par sommes et produits de ses coefficients ; réduire tous les coefficients modulo m ne change donc pas son reste modulo m. Ainsi, $\det(A)\det(B) = \det(AB) \equiv \det(I) = 1 \pmod{m}$, et $\det A$ est inversible modulo m, ce qui exige $\operatorname{pgcd}(\det A, m) = 1$ (proposition E.19). □

Pour une matrice 2×2, la matrice adjointe s'obtient en échangeant les coefficients de la diagonale et en changeant le signe des deux autres :

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \qquad \operatorname{adj}(A) = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}, \qquad A^{-1} \equiv \delta \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \pmod{26},$$

avec $\delta(ad - bc) \equiv 1 \pmod{26}$. C'est la formule bien connue de l'inverse d'une matrice 2×2, dans laquelle on a remplacé la division par $ad - bc$ par une multiplication par son inverse modulaire.

Exemple 7.10 **Déchiffrer OZYRFRAG**

Pour la matrice de l'exemple 7.9, $\det A = 21 - 10 = 11$, et $\operatorname{pgcd}(11, 26) = 1$. L'inverse de 11 modulo 26 est 19, car $11 \times 19 = 209 = 8 \times 26 + 1$. Ainsi,

$$A^{-1} \equiv 19 \begin{pmatrix} 7 & -2 \\ -5 & 3 \end{pmatrix} = \begin{pmatrix} 133 & -38 \\ -95 & 57 \end{pmatrix} \equiv \begin{pmatrix} 3 & 14 \\ 9 & 5 \end{pmatrix} \pmod{26}.$$

On vérifie que $A^{-1}A = \begin{pmatrix} 9+70 & 6+98 \\ 27+25 & 18+35 \end{pmatrix} = \begin{pmatrix} 79 & 104 \\ 52 & 53 \end{pmatrix} \equiv I \pmod{26}$. Le bloc OZ, soit $(14, 25)$, donne $(3 \times 14 + 14 \times 25,\ 9 \times 14 + 5 \times 25) = (392, 251) \equiv (2, 17)$, c'est-à-dire CR.

Combien de clés le chiffre de Hill offre-t-il ? Pour $n = 2$, il y a $26^4 = 456\,976$ matrices, dont 157 248 sont inversibles modulo 26 (exercice 7.8) ; pour $n = 3$, il y en a plus de $1{,}6 \times 10^{12}$. Le nombre de clés croît très vite avec la taille des blocs.

Le programme 7.3 chiffre avec une matrice de taille quelconque. Le déterminant et les cofacteurs sont calculés avec des entiers exacts, par développement selon la première ligne ; on évite ainsi les erreurs d'arrondi des fonctions de `numpy`, qui travaillent avec des nombres à virgule.

Programme 7.3 : Le chiffre de Hill (on utilise les fonctions du programme 7.1).

```
def produit_mod(M, v, m=26):
    return [sum(M[i][k] * v[k] for k in range(len(v))) % m for i in
    range(len(M))]

def determinant(M):
    """Déterminant exact, par développement selon la première ligne."""
    if len(M) == 1:
        return M[0][0]
    return sum((-1) ** j * M[0][j] * determinant([ligne[:j] + ligne[j
    +1:] for ligne in M[1:]])
```

```
                for j in range(len(M)))

def inverse_mod(M, m=26):
    n = len(M)
    delta = pow(determinant(M) % m, -1, m)   # erreur si pgcd(det, m) !=
     1
    def cofacteur(i, j):
        mineur = [ligne[:j] + ligne[j+1:] for k, ligne in enumerate(M)
    if k != i]
        return (-1) ** (i + j) * determinant(mineur)
    # inverse = delta x adj(M), où adj(M) est la transposée des
    cofacteurs
    return [[delta * cofacteur(j, i) % m for j in range(n)] for i in
    range(n)]

def hill(texte, M):
    n = len(M)
    x = vers_nombres(texte)
    x += [23] * (-len(x) % n)                # on complète avec des x
    y = []
    for i in range(0, len(x), n):
        y += produit_mod(M, x[i:i+n])
    return vers_lettres(y)

A = [[3, 2], [5, 7]]
print(hill("cryptage", A))                   # OZYRFRAG
print(inverse_mod(A))                        # [[3, 14], [9, 5]]
print(hill("OZYRFRAG", inverse_mod(A)))      # CRYPTAGE
```

7.6.3 Une attaque à texte clair connu

Le chiffre de Hill résiste à l'analyse des fréquences des lettres, mais il a un défaut grave : il est *linéaire*. Si Ève connaît n blocs clairs et les blocs chiffrés correspondants, elle peut les placer en colonnes dans deux matrices X et Y, de sorte que $Y \equiv AX$. Si X est inversible modulo 26, elle obtient immédiatement la clé :

$$A \equiv Y X^{-1} \pmod{26}.$$

Exemple 7.11 **Retrouver la matrice**

Ève a intercepté `OZYRFRAG` et sait que le message est le mot `cryptage`. Les blocs `cr` et `ta` sont devenus `OZ` et `FR`. Elle forme

$$X = \begin{pmatrix} 2 & 19 \\ 17 & 0 \end{pmatrix}, \qquad Y = \begin{pmatrix} 14 & 5 \\ 25 & 17 \end{pmatrix}.$$

On a $\det X = -323 \equiv 15 \pmod{26}$, dont l'inverse est 7 ($15 \times 7 = 105 \equiv 1$). Ainsi

$$X^{-1} \equiv 7 \begin{pmatrix} 0 & -19 \\ -17 & 2 \end{pmatrix} \equiv \begin{pmatrix} 0 & 23 \\ 11 & 14 \end{pmatrix}, \qquad A \equiv YX^{-1} = \begin{pmatrix} 55 & 392 \\ 187 & 813 \end{pmatrix} \equiv \begin{pmatrix} 3 & 2 \\ 5 & 7 \end{pmatrix} \pmod{26}.$$

Elle a retrouvé la clé, et peut lire tous les autres messages chiffrés avec elle. Si elle avait choisi les blocs `cr` et `yp`, la matrice X aurait eu pour déterminant $30 - 408 = -378 \equiv 12$, qui n'est pas inversible modulo 26 : il aurait fallu utiliser un autre couple de blocs.

Cette faiblesse est commune à tous les chiffrements linéaires. Les chiffrements modernes à clé secrète, comme l'AES, adopté en 2001 et utilisé aujourd'hui pour chiffrer la plupart des données sur Internet, reprennent l'idée de Hill (des blocs, des opérations algébriques, de la diffusion), mais ils y ajoutent une étape *non linéaire* et répètent l'ensemble une dizaine de fois.

Pour la classe 7.4 **Des matrices sans le dire**

Les matrices ne figurent au programme que de la séquence Technico-sciences, où elles servent surtout à organiser des données [Min16]. Le chiffre de Hill 2×2 peut toutefois se présenter sans le vocabulaire matriciel : chaque paire de lettres (x_1, x_2) devient $(3x_1 + 2x_2,\ 5x_1 + 7x_2)$, réduit modulo 26. Les élèves chiffrent sans difficulté. Le déchiffrement est un système de deux équations à deux inconnues, que l'on résout par élimination, et la condition $\operatorname{pgcd}(\det A, 26) = 1$ apparaît naturellement au moment de « diviser ». C'est une belle occasion de montrer que les mêmes techniques algébriques fonctionnent dans d'autres systèmes de nombres que $\mathbb{R}$, pourvu qu'on sache y faire les quatre opérations.

7.7 Enigma, la mécanisation du chiffrement

7.7.1 La machine

Au début du XXe siècle, la radio transforme la guerre : les ordres voyagent instantanément, mais l'ennemi peut les capter. Il faut chiffrer vite, beaucoup et sans erreur, ce que le crayon et le papier ne permettent plus. L'ingénieur allemand Arthur Scherbius dépose en 1918 le brevet d'une machine électromécanique, l'Enigma, d'abord vendue aux entreprises, puis adoptée par l'armée allemande à la fin des années 1920. Pendant la Seconde Guerre mondiale, des dizaines de milliers d'exemplaires chiffrent les communications de l'armée de terre, de l'aviation et de la marine allemandes.

Vue de l'extérieur, l'Enigma ressemble à une machine à écrire : un clavier, et au-dessus, un tableau de 26 lampes. Lorsque l'opérateur enfonce une touche, une lampe s'allume,

qui donne la lettre chiffrée. Entre les deux, le courant traverse les pièces représentées à la figure 7.4 :

— le **tableau de connexions**, où des câbles échangent des paires de lettres (dix paires, en général) ;
— trois **rotors**, des disques dont chaque face porte 26 contacts, reliés deux à deux par un câblage interne fixe : chaque rotor réalise une permutation des 26 lettres ;
— le **réflecteur**, qui relie les contacts deux à deux et renvoie le courant à travers les rotors, par un autre chemin, jusqu'au tableau de connexions puis à une lampe.

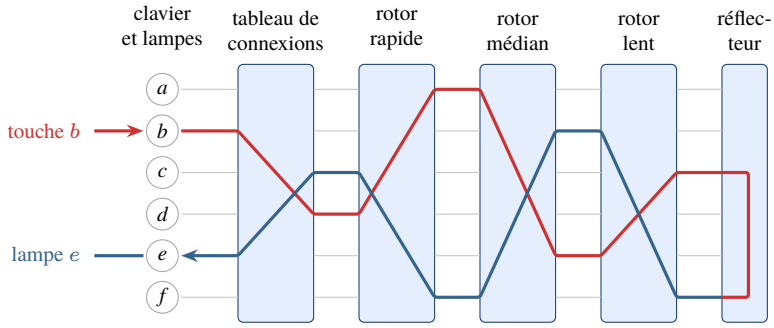


FIGURE 7.4 : Le principe de l'Enigma, illustré avec six lettres au lieu de vingt-six. Le courant de la touche b traverse le tableau de connexions, les trois rotors et le réflecteur (en rouge), puis revient par un autre chemin (en bleu) et allume la lampe e. Après chaque frappe, le rotor rapide tourne d'un cran, et tous les câblages sont décalés.

Ce qui fait la force de l'Enigma, c'est que les rotors *tournent*. Après chaque lettre, le rotor rapide avance d'un cran ; après un tour complet, il entraîne le rotor médian d'un cran, à la manière d'un compteur kilométrique, et le rotor médian entraîne de même le rotor lent. La substitution change donc à chaque lettre : l'Enigma est un chiffre polyalphabétique, comme celui de Vigenère, mais dont la « clé » ne se répète qu'après près de 17 000 lettres, bien plus que la longueur d'un message.

7.7.2 Une involution sans point fixe

Notons P la permutation réalisée par le tableau de connexions, R la permutation réalisée par les trois rotors dans leur position du moment (pour le trajet aller), et U celle du réflecteur. Le trajet retour traverse les mêmes pièces en sens inverse, et réalise donc les permutations réciproques. La lettre x devient ainsi

$$E(x) = P^{-1}(R^{-1}(U(R(P(x))))), \qquad \text{soit} \qquad E = \sigma^{-1} \circ U \circ \sigma, \quad \text{où } \sigma = R \circ P.$$

Le réflecteur relie les contacts deux à deux : U échange les lettres par paires, de sorte que $U \circ U = \mathrm{id}$ et que $U(z) \neq z$ pour toute lettre z.

Proposition 7.3 **Les deux propriétés de l'Enigma**

Pour chaque position des rotors, la permutation E réalisée par l'Enigma vérifie :

(i) $E \circ E = \mathrm{id}$: déchiffrer, c'est chiffrer une seconde fois, avec le même réglage ;

(ii) $E(x) \neq x$ pour toute lettre x : une lettre n'est jamais chiffrée par elle-même.

Démonstration. (i) On a $E \circ E = \sigma^{-1} \circ U \circ \sigma \circ \sigma^{-1} \circ U \circ \sigma = \sigma^{-1} \circ U \circ U \circ \sigma = \sigma^{-1} \circ \sigma = \mathrm{id}$.

(ii) Si $E(x) = x$, alors $U(\sigma(x)) = \sigma(x)$, en appliquant σ aux deux membres. La lettre $z = \sigma(x)$ serait un point fixe de U, ce qui est impossible.

□

La propriété (i) était voulue : la même machine, réglée de la même façon, sert à chiffrer et à déchiffrer, ce qui simplifie l'usage. La propriété (ii), qui en est le prix, s'est révélée être une faiblesse fatale. Supposons que les cryptanalystes devinent un mot du message, par exemple `wetterbericht` (« bulletin météo »), présent chaque matin dans certains messages. Ils ne savent pas où il se trouve, mais ils peuvent éliminer toutes les positions où une lettre du mot coïncide avec la lettre chiffrée correspondante. Ce genre de mot probable, appelé *crib* par les Britanniques, était le point de départ de leurs attaques.

7.7.3 Le nombre de clés

Combien de réglages l'opérateur doit-il choisir ? Comptons, pour la version la plus répandue dans l'armée de terre allemande.

- L'*ordre des rotors* : on en place trois, choisis parmi cinq et rangés dans un certain ordre, soit $5 \times 4 \times 3 = 60$ possibilités.
- La *position de départ* des trois rotors : $26^3 = 17\,576$ possibilités.
- Le *tableau de connexions* : on relie dix paires de lettres. On choisit d'abord les 20 lettres reliées, puis on les regroupe en dix paires. En choisissant les paires une à une, on obtient $\binom{26}{2}\binom{24}{2}\cdots\binom{8}{2}$; mais chaque ensemble de dix paires a alors été compté 10! fois, une fois pour chaque ordre dans lequel on peut choisir les mêmes paires. Il y a donc

$$\frac{1}{10!}\binom{26}{2}\binom{24}{2}\cdots\binom{8}{2} = \frac{26!}{6!\,10!\,2^{10}} = 150\,738\,274\,937\,250 \approx 1{,}5 \times 10^{14}$$

  façons de brancher le tableau.

Au total, par le principe multiplicatif, il y a environ

$$60 \times 17\,576 \times 1{,}5 \times 10^{14} \approx 1{,}6 \times 10^{20}$$

réglages, sans compter les bagues de chaque rotor, qui ajoutaient encore un facteur. Une attaque exhaustive était hors de question avec les moyens de l'époque. Remarquons que le tableau de connexions, qui ne semble être qu'un accessoire, fournit à lui seul la plus grande partie de ces clés.

Note historique 7.2 Casser Enigma : de Varsovie à Bletchley Park

Les premiers à casser l'Enigma militaire furent trois jeunes mathématiciens polonais, Marian Rejewski, Jerzy Różycki et Henryk Zygalski, recrutés par le service du chiffre polonais en 1932. Rejewski reconstitua le câblage des rotors, sans avoir jamais vu la machine, en appliquant la théorie des permutations, alors fort peu utilisée en dehors des mathématiques, aux messages interceptés et à des documents fournis par un espion. Les Polonais construisirent des machines, appelées *bomby*, pour retrouver chaque jour les réglages. En juillet 1939, quelques semaines avant l'invasion de leur pays, ils transmirent tous leurs résultats aux services français et britanniques.

Les Britanniques réunirent à Bletchley Park, au nord de Londres, des mathématiciens, des linguistes, des champions d'échecs et des cruciverbistes. Alan Turing y conçut, avec Gordon Welchman, la *bombe*, une machine électromécanique qui testait rapidement les réglages compatibles avec un *crib*, en exploitant notamment la proposition 7.3. La capture de documents de chiffrement, comme ceux du sous-marin U-110 en mai 1941, aida aussi à percer les réglages de la marine. Pour casser

un autre chiffrement allemand, celui de la machine Lorenz, l'ingénieur Tommy Flowers construisit en 1944 le *Colossus*, l'un des premiers ordinateurs électroniques programmables. Le secret de Bletchley Park fut gardé jusqu'aux années 1970 [Sin99].

Outil 7.2 **Simuler une Enigma**

Plusieurs simulateurs d'Enigma sont accessibles dans un navigateur, notamment sur *CrypTool-Online*. On y vérifie en quelques frappes les deux propriétés de la proposition 7.3. L'exercice 7.17 propose de programmer une Enigma simplifiée en Python.

Pour la classe 7.5 **Dénombrer les clés**

Le dénombrement des clés de l'Enigma, comme celui des substitutions (26!) ou des clés du chiffre affine (12×26), est un excellent problème de dénombrement pour le secondaire. Le principe multiplicatif suffit pour les rotors, et le calcul des branchements du tableau de connexions fait apparaître une difficulté classique : éviter de compter plusieurs fois la même configuration. On peut commencer par une version réduite, avec 6 lettres et 2 câbles, que les élèves énumèrent à la main (il y a 45 branchements), avant de généraliser.

7.8 La cryptographie à clé publique

7.8.1 Le problème de l'échange des clés

Tous les chiffrements rencontrés jusqu'ici sont **symétriques** : la même clé sert à chiffrer et à déchiffrer, ou du moins chacune se déduit facilement de l'autre. Alice et Bob doivent donc s'être mis d'accord sur une clé *avant* de communiquer, par un moyen sûr : une rencontre, un messager, une valise diplomatique. Pour une armée, c'est une lourde logistique ; pour Internet, où l'on communique à chaque instant avec des serveurs que l'on n'a jamais rencontrés, c'est impossible. Avec N personnes qui veulent toutes communiquer deux à deux, il faudrait d'ailleurs $\binom{N}{2} = N(N-1)/2$ clés différentes.

Pendant des siècles, ce problème a semblé sans solution : comment partager un secret sans s'être d'abord partagé un secret ? Une image permet de comprendre que c'est possible. Bob distribue à tout le monde des *cadenas ouverts*, dont il est le seul à posséder la clé. Pour lui écrire, Alice met son message dans une boîte, la ferme avec un cadenas de Bob et l'envoie. Ève peut intercepter la boîte, mais pas l'ouvrir ; Alice elle-même ne pourrait plus l'ouvrir. Fermer le cadenas est à la portée de tous, l'ouvrir est réservé à Bob.

En mathématiques, le cadenas est une fonction facile à calculer, mais difficile à inverser, sauf pour qui détient un renseignement supplémentaire. C'est l'idée des **systèmes à clé publique**, publiée en 1976 par Whitfield Diffie et Martin Hellman, de l'Université Stanford [DH76]. On sait aujourd'hui que des chercheurs des services secrets britanniques, James Ellis, Clifford Cocks et Malcolm Williamson, avaient fait des découvertes équivalentes quelques années plus tôt, mais elles sont restées secrètes jusqu'en 1997.

7.8.2 L'échange de clés de Diffie et Hellman

Diffie et Hellman n'ont pas proposé un chiffrement, mais une façon, pour Alice et Bob, de fabriquer une clé secrète commune en ne s'échangeant que des messages publics. Leur fonction « facile à calculer, difficile à inverser » est l'exponentiation modulaire.

Algorithme 7.1 **Échange de clés de Diffie et Hellman**

Alice et Bob conviennent publiquement d'un grand nombre premier p et d'un entier g, avec $1 < g < p$.

1. Alice choisit un entier secret a et envoie à Bob $A = g^a \bmod p$.
2. Bob choisit un entier secret b et envoie à Alice $B = g^b \bmod p$.
3. Alice calcule $B^a \bmod p$, et Bob calcule $A^b \bmod p$.

Les deux résultats sont égaux, puisque $B^a \equiv (g^b)^a = g^{ab} = (g^a)^b \equiv A^b \pmod p$. Alice et Bob partagent donc le nombre $s = g^{ab} \bmod p$, qu'ils peuvent utiliser comme clé d'un chiffrement symétrique.

Exemple 7.12 **Diffie et Hellman avec de petits nombres**

Prenons $p = 29$ et $g = 2$. Alice choisit $a = 5$ et envoie $A = 2^5 = 32 \equiv 3 \pmod{29}$. Bob choisit $b = 12$ et envoie $B = 2^{12} = 4096 = 141 \times 29 + 7 \equiv 7 \pmod{29}$. Alice calcule $7^5 = 16\,807 \equiv 16 \pmod{29}$, et Bob calcule $3^{12} = 531\,441 \equiv 16 \pmod{29}$. La clé commune est $s = 16$.

Ève a vu passer $p = 29$, $g = 2$, $A = 3$ et $B = 7$. Pour trouver s, il lui suffirait de connaître a, c'est-à-dire de résoudre $2^a \equiv 3 \pmod{29}$. Ce problème, appelé le **logarithme discret**, se résout ici en essayant les exposants un à un. Mais pour un nombre premier p de plusieurs centaines de chiffres, on ne connaît aucune méthode beaucoup plus rapide que l'essai d'un nombre astronomique de possibilités, alors que le calcul de $g^a \bmod p$ ne prend qu'une fraction de seconde (section 7.8.4). Le logarithme ordinaire, que les élèves rencontrent au deuxième cycle du secondaire, se calcule facilement parce que la fonction exponentielle est croissante ; modulo p, les puissances de g « sautent » d'un bout à l'autre de $\{1, \ldots, p-1\}$ sans aucune régularité apparente.

7.8.3 Le chiffrement RSA

Un an plus tard, en 1977, trois chercheurs du MIT, Ronald Rivest, Adi Shamir et Leonard Adleman, proposèrent le premier véritable chiffrement à clé publique, connu depuis sous le nom de RSA [RSA78]. Sa fonction « cadenas » est l'élévation à une puissance modulo un nombre n, et le renseignement secret qui permet d'ouvrir le cadenas est la factorisation de n. Nous en avons donné le principe à la remarque E.9 ; reprenons-le en détail.

Algorithme 7.2 **Le chiffrement RSA**

Création des clés, par Bob.

1. Choisir deux grands nombres premiers distincts p et q, et calculer $n = pq$ et $\varphi(n) = (p-1)(q-1)$.
2. Choisir un entier e, avec $1 < e < \varphi(n)$ et $\text{pgcd}(e, \varphi(n)) = 1$.
3. Calculer l'inverse d de e modulo $\varphi(n)$, à l'aide de l'algorithme d'Euclide étendu (section E.4).
4. Publier la **clé publique** (n, e), et garder secrète la **clé privée** d (ainsi que p, q et $\varphi(n)$).

Chiffrement, par Alice. Le message est un entier m avec $0 \leq m < n$. Alice envoie $c = m^e \mod n$.
Déchiffrement, par Bob. Bob calcule $c^d \mod n$, qui est égal à m.

La formule $\varphi(n) = (p-1)(q-1)$ découle de la proposition E.23 : parmi les entiers de 1 à pq, ceux qui ne sont pas premiers avec pq sont les q multiples de p et les p multiples de q, pq étant compté deux fois, d'où $\varphi(pq) = pq - q - p + 1 = (p-1)(q-1)$. Le théorème suivant justifie le déchiffrement, pour *tout* message m.

Théorème 7.4 **Le déchiffrement RSA fonctionne**

Soient $p \neq q$ deux nombres premiers, $n = pq$, et e, d deux entiers positifs tels que $ed \equiv 1 \pmod{\varphi(n)}$. Alors, pour tout entier m,

$$(m^e)^d = m^{ed} \equiv m \pmod{n}.$$

Démonstration. Écrivons $ed = 1 + k(p-1)(q-1)$, avec $k \geq 0$ un entier. Montrons d'abord que p divise $m^{ed} - m$. Si p divise m, il divise aussi m^{ed}, et donc leur différence. Sinon, le petit théorème de Fermat (théorème E.21) donne $m^{p-1} \equiv 1 \pmod{p}$, d'où

$$m^{ed} = m \cdot \left(m^{p-1}\right)^{k(q-1)} \equiv m \cdot 1 = m \pmod{p}.$$

De la même façon, q divise $m^{ed} - m$. Comme p et q sont des nombres premiers distincts, ils sont premiers entre eux, et leur produit $n = pq$ divise $m^{ed} - m$ (corollaire E.13). □

Lorsque $\text{pgcd}(m, n) = 1$, on peut aussi invoquer directement le théorème d'Euler, comme à la remarque E.9 : $m^{ed} = m \cdot (m^{\varphi(n)})^k \equiv m$. La démonstration ci-dessus a l'avantage de couvrir aussi les messages divisibles par p ou par q (qui sont, de toute façon, extrêmement rares lorsque p et q sont grands).

Exemple 7.13 **RSA avec de petits nombres premiers**

Bob choisit $p = 47$ et $q = 59$. Il calcule $n = 2773$ et $\varphi(n) = 46 \times 58 = 2668$. Il choisit $e = 17$, qui est premier avec $2668 = 2^2 \times 23 \times 29$. L'algorithme d'Euclide donne

$$2668 = 156 \times 17 + 16, \qquad 17 = 1 \times 16 + 1,$$

d'où $1 = 17 - 16 = 17 - (2668 - 156 \times 17) = 157 \times 17 - 2668$. L'inverse de 17 modulo 2668 est donc $d = 157$. Bob publie $(n, e) = (2773, 17)$ et garde $d = 157$ secret.
Alice veut envoyer `bonjour`. Elle découpe le message en blocs de deux lettres, `bo`, `nj`, `ou` et `rx` (elle complète avec un x), et écrit chaque bloc comme un nombre de quatre chiffres : `bo` devient 0114, puisque $b = 01$ et $o = 14$. Comme le plus grand bloc possible, `zz`, donne $2525 < n$, chaque bloc est un message admissible. Elle obtient

$$m = 0114,\ 1309,\ 1420,\ 1723 \quad \longmapsto \quad c = m^{17} \bmod 2773 = 1684,\ 0840,\ 0381,\ 1339.$$

Bob calcule, par exemple, $1684^{157} \bmod 2773 = 114$, et retrouve `bo`.

On remarque que, contrairement aux chiffrements des sections précédentes, *n'importe qui* peut chiffrer un message pour Bob : la clé de chiffrement est publique. La sécurité repose entièrement sur la difficulté de retrouver d à partir de (n, e).

7.8.4 L'exponentiation rapide

Calculer 1684^{157} puis réduire modulo 2773 serait absurde : ce nombre a plus de 500 chiffres. Dans la vraie vie, avec des exposants de plus de 600 chiffres, il en aurait davantage qu'il n'y a d'atomes dans l'Univers. Deux idées rendent le calcul très rapide. D'abord, on réduit modulo n après *chaque* multiplication (proposition E.18) : les nombres manipulés restent inférieurs à n^2. Ensuite, on utilise les carrés successifs $a, a^2, a^4, a^8, \ldots$, obtenus en élevant au carré le précédent, et l'écriture de l'exposant en base deux (section 6.1).

Exemple 7.14 **Calculer $114^{17} \bmod 2773$**

Comme $17 = (10001)_2 = 16 + 1$, on a $114^{17} = 114^{16} \cdot 114$. On élève au carré quatre fois, en réduisant à chaque étape :

$$114^2 \equiv 1904, \quad 114^4 \equiv 1904^2 \equiv 905, \quad 114^8 \equiv 905^2 \equiv 990, \quad 114^{16} \equiv 990^2 \equiv 1231,$$

puis $114^{17} \equiv 1231 \times 114 = 140\,334 \equiv 1684 \pmod{2773}$. Cinq multiplications ont suffi, au lieu de seize.

En général, le nombre de carrés à calculer est le nombre de chiffres de l'exposant k en base deux, moins un, soit $\lfloor \log_2 k \rfloor$, et il faut au plus autant de multiplications supplémentaires. Pour un exposant de 2048 bits, il faut donc au plus 4096 multiplications, ce qu'un téléphone fait en un instant. La fonction `pow(a, k, n)` de Python applique cette méthode.

Programme 7.4 : Le chiffrement RSA de l'exemple 7.13 (on utilise la fonction `vers_nombres` du programme 7.1).

```
def puissance_mod(a, k, n):
    """Exponentiation rapide : renvoie a^k mod n."""
    r = 1
    a = a % n
```

```
    while k > 0:
        if k % 2 == 1:
            r = r * a % n
        a = a * a % n
        k = k // 2
    return r

p, q, e = 47, 59, 17
n, phi = p * q, (p - 1) * (q - 1)
d = pow(e, -1, phi)                        # clé privée : 157

def blocs(texte):
    """Blocs de deux lettres, écrits comme des nombres de 4 chiffres."""
    x = vers_nombres(texte)
    x += [23] * (len(x) % 2)               # on complète avec un x
    return [100 * x[i] + x[i + 1] for i in range(0, len(x), 2)]

m = blocs("bonjour")                       # [114, 1309, 1420, 1723]
c = [puissance_mod(b, e, n) for b in m]    # [1684, 840, 381, 1339]
print(m, c, [pow(b, d, n) for b in c])
```

7.8.5 Pourquoi RSA est-il sûr ?

Ève connaît n et e. Pour calculer d, il lui suffirait de connaître $\varphi(n)$, et donc p et q : elle doit *factoriser* n. Avec $n = 2773$, c'est immédiat : il suffit d'essayer de diviser par les nombres premiers inférieurs à $\sqrt{2773} \approx 52{,}7$, et l'on trouve $2773 = 47 \times 59$. Mais si p et q ont chacun plus de 300 chiffres, le nombre de divisions à essayer dépasse de très loin les capacités de tous les ordinateurs du monde, et les meilleures méthodes connues, beaucoup plus subtiles, restent elles aussi hors d'atteinte. La sécurité de RSA repose sur une asymétrie remarquable : *multiplier* deux grands nombres premiers est facile, retrouver les facteurs à partir du produit est extraordinairement difficile.

Cette difficulté n'est pas démontrée ; c'est un constat, sans cesse remis à l'épreuve. En 1977, Martin Gardner publia dans le *Scientific American* un défi des auteurs de RSA : un message chiffré avec un nombre n de 129 chiffres. Il fallut attendre 1994 et le travail coordonné de centaines d'ordinateurs pour le factoriser. Le record actuel, établi en 2020, est la factorisation d'un nombre de 250 chiffres (829 bits). C'est pourquoi on recommande aujourd'hui des clés d'au moins 2048 bits, soit environ 617 chiffres. Par ailleurs, Peter Shor a montré en 1994 qu'un ordinateur *quantique* suffisamment puissant pourrait factoriser rapidement de grands nombres, et calculer aussi des logarithmes discrets. De telles machines n'existent pas encore, mais, par précaution, de nouveaux algorithmes « post-quantiques », fondés sur d'autres problèmes mathématiques, ont été normalisés en 2024 et commencent à remplacer RSA et Diffie–Hellman.

En pratique, RSA est lent comparé aux chiffrements symétriques. Lors d'une connexion sécurisée à un site Web (le petit cadenas du navigateur), on utilise la cryptographie à clé publique pour s'authentifier et pour convenir d'une clé secrète, puis un chiffrement symétrique rapide, comme l'AES, pour échanger les données. Les deux familles de chiffrements sont donc complémentaires.

7.8.6 La signature numérique

La cryptographie à clé publique offre un autre service, peut-être plus précieux encore que la confidentialité : l'*authentification*. Comment Bob peut-il être certain qu'un message vient bien d'Alice, et qu'il n'a pas été modifié ? Il suffit d'inverser les rôles des deux clés.

Alice, dont la clé publique est (n_A, e_A) et la clé privée d_A, joint à son message m la **signature**

$$s = m^{d_A} \bmod n_A,$$

qu'elle seule peut calculer. N'importe qui peut la vérifier, à l'aide de la clé publique d'Alice : par le théorème 7.4, $s^{e_A} = m^{d_A e_A} \equiv m \pmod{n_A}$. Si le message a été modifié, ou si la signature a été fabriquée par quelqu'un d'autre, l'égalité n'est pas vérifiée. Pour garder le message confidentiel, Alice peut ensuite chiffrer le tout avec la clé publique de Bob.

Les signatures numériques garantissent l'authenticité des mises à jour de logiciels, des certificats des sites Web et des transactions bancaires.

Pour la classe 7.6 **RSA en classe**

> Le chiffrement RSA peut se faire en classe avec de petits nombres et une calculatrice. Avec $p = 5$ et $q = 11$, on a $n = 55$, $\varphi(n) = 40$, $e = 3$ et $d = 27$; le message $m = 8$ devient $8^3 = 512 \equiv 17 \pmod{55}$. Le déchiffrement, $17^{27} \bmod 55$, est une belle occasion de réinvestir les lois des exposants : $17^{27} = 17^{16} \cdot 17^8 \cdot 17^2 \cdot 17$, chaque facteur s'obtenant du précédent par une élévation au carré. Les élèves y découvrent qu'un calcul impossible à la calculatrice devient facile lorsqu'on réduit à chaque étape. On peut aussi organiser un « annuaire » de clés publiques de la classe : chaque équipe publie sa clé, et les autres lui envoient des messages qu'elle seule peut lire, puis tentent de casser les clés des autres en factorisant n.

Exercices

Exercice 7.1. On utilise le chiffrement par décalage (définition 7.2).

a) Chiffrer le mot `guitare` avec le chiffre de César.
b) Le mot `DXWR` a été chiffré avec le chiffre de César. Quel était le texte clair ?
c) Un chiffrement par décalage transforme `chien` en `HMNJS`. Comment chiffre-t-il le mot `chat` ?
d) Combien de fois faut-il appliquer le chiffre de César à un texte pour retrouver le texte de départ ? Et un décalage de clé k quelconque ?

Exercice 7.2. On utilise le chiffre de Vigenère (définition 7.4).

a) Chiffrer `jaimelesexpos` avec la clé $(1, 3, 24)$.
b) Montrer qu'un chiffre de Vigenère dont la clé est de longueur 1 est un chiffrement par décalage.
c) On chiffre un texte avec une clé de longueur ℓ_1, puis on chiffre le résultat avec une clé de longueur ℓ_2. Montrer que le résultat est un chiffre de Vigenère dont la clé est de longueur au plus $\text{ppcm}(\ell_1, \ell_2)$. Est-ce une bonne idée pour renforcer le chiffrement ?

Exercice 7.3. On utilise le chiffre affine de clé $(\alpha, \beta) = (5, 8)$.

a) Chiffrer le mot `secondaire`.
b) Trouver la clé de déchiffrement, puis déchiffrer `FYZRIMAPC`.
c) Une clé affine est une *involution* si chiffrer deux fois redonne le texte clair. Trouver toutes les clés affines qui sont des involutions.

Exercice 7.4. Ève intercepte le message `MUZMCK`, chiffré par un chiffre affine. Elle sait que le message commence par les lettres `eu`. Retrouver la clé et le message.

Exercice 7.5. Une enseignante a chiffré les prénoms de six élèves, Camille, Félix, Noémie, Olivier, Rosalie et Samuel, avec un chiffre affine, après avoir retiré les accents. Elle a ensuite mélangé la liste chiffrée :

`MFXHYRZ`, `KZYRA`, `XHJTZY`, `DHJRYYZ`, `FYRERZM`, `UFZJRZ`.

a) Associer chaque prénom à son chiffré.
b) Trouver la clé de chiffrement, et la clé de déchiffrement.

Exercice 7.6. Cet exercice reprend une question d'un travail pratique sur les substitutions monoalphabétiques. On note $E = \{a, b, \dots, z\}$, et l'on appelle **permutation** de E toute bijection de E dans E. Voici un exemple de permutation :

$$\begin{pmatrix} a & b & c & d & e & f & g & h & i & j & k & l & m & n & o & p & q & r & s & t & u & v & w & x & y & z \\ g & f & e & d & c & b & a & h & i & j & z & x & y & k & l & m & q & r & s & t & u & v & w & p & o & n \end{pmatrix}.$$

a) Combien y a-t-il de permutations de E ?
b) Un chiffre affine de clé admissible définit-il toujours une permutation de E ?
c) Toute permutation de E est-elle un chiffre affine ? La permutation ci-dessus en est-elle un ?
d) Une lettre peut-elle être inchangée par un chiffre affine ? Décrire les lettres inchangées par le chiffre affine de clé (α, β).

Exercice 7.7. On utilise le chiffre de Hill de clé $K = \begin{pmatrix} 4 & 1 \\ -1 & 2 \end{pmatrix}$.

a) Calculer l'inverse de K dans $\mathbb{R}$, puis montrer que K est inversible modulo 26 et calculer son inverse modulo 26.
b) Déchiffrer `WOFVCUTMWYGORC`.
c) Ève sait que les blocs `th` et `at` du texte clair sont devenus `FV` et `TM`. Montrer qu'elle peut retrouver K.
d) Ève sait aussi que `ma` est devenu `WO`. Pourquoi les blocs `ma` et `th` ne lui suffiraient-ils pas ?

Exercice 7.8. On veut compter les clés du chiffre de Hill 2×2, c'est-à-dire les matrices 2×2 à coefficients dans $\mathbb{Z}_{26}$ qui sont inversibles modulo 26.

a) Soit p un nombre premier. Montrer qu'une matrice 2×2 à coefficients dans $\mathbb{Z}_p$ est inversible modulo p si et seulement si sa première colonne est non nulle et sa seconde colonne n'est pas un multiple de la première. En déduire qu'il y a $(p^2-1)(p^2-p)$ matrices inversibles modulo p.
b) À l'aide du théorème des restes chinois (théorème E.25) appliqué à chaque coefficient, montrer qu'une matrice est inversible modulo 26 si et seulement si elle est inversible modulo 2 et modulo 13, et en déduire le nombre de clés.

Exercice 7.9. On reprend le dénombrement des branchements du tableau de connexions de l'Enigma (section 7.7).

a) Avec 6 lettres et 2 câbles, combien y a-t-il de branchements ? Les énumérer pour vérifier.

b) Montrer qu'avec 26 lettres et k câbles ($0 \leq k \leq 13$), il y a

$$N_k = \frac{26!}{(26-2k)!\; k!\; 2^k}$$

branchements.

c) À l'aide d'une calculatrice ou de Python, calculer N_k pour $k = 9, 10, 11, 12, 13$. Quel nombre de câbles donne le plus de clés ? Pourquoi utiliser tous les câbles n'est-il pas le meilleur choix ?

Exercice 7.10. Bob choisit $p = 7$, $q = 17$ et $e = 5$.

a) Calculer n et $\varphi(n)$, vérifier que e est admissible et calculer la clé privée d à l'aide de l'algorithme d'Euclide étendu.

b) Alice envoie le message $m = 10$. Calculer le message chiffré c.

c) Déchiffrer c à l'aide de l'exponentiation rapide, sans calculatrice.

d) Ève connaît la clé publique (n, e). Comment peut-elle retrouver d ?

Exercice 7.11. Wilbrod utilise RSA, avec la clé publique (n, e). Pauline, de nature méfiante, chiffre son message m (premier avec n) pour obtenir c, puis chiffre c une seconde fois avec la même clé, et envoie le résultat b à Wilbrod. Pour son plus grand malheur, la clé publique de Wilbrod vérifie $e^2 \equiv 1 \pmod{\varphi(n)}$.

a) Expliquer pourquoi toute personne qui intercepte b lit le message de Pauline sans effort.

b) Si Wilbrod avait choisi $p = 3$ et $q = 11$, quelles valeurs de e aurait-il dû éviter ?

Exercice 7.12. Maria et Gontrand utilisent RSA avec le même module n, mais avec des exposants différents : leurs clés publiques sont (n, e_M) et (n, e_G), avec $\operatorname{pgcd}(e_M, e_G) = 1$. Édouardo leur envoie le même message m, premier avec n ; il transmet $c_M = m^{e_M} \bmod n$ à Maria et $c_G = m^{e_G} \bmod n$ à Gontrand. Dimitri intercepte les deux messages chiffrés.

a) Montrer que Dimitri peut retrouver m sans connaître aucune des clés privées.

b) Si $e_M = 10$ et $e_G = 7$, trouver deux entiers x et y tels que $e_M x + e_G y = 1$, et donner la formule que Dimitri utilise.

Exercice 7.13. On choisit deux nombres premiers p et q congrus à 2 modulo 3, et l'on pose $n = pq$. Le message est un entier $x \in \{1, \ldots, n-1\}$ premier avec n. On le chiffre en calculant $y = x^3 \bmod n$, et le destinataire déchiffre en calculant $z = y^d \bmod n$, où

$$d = \frac{2(p-1)(q-1)+1}{3}.$$

a) Montrer que d est un entier.

b) Montrer que y et z ne sont pas congrus à 0 modulo n.

c) Montrer que $z \equiv x \pmod{n}$, c'est-à-dire que le déchiffrement fonctionne. Quel est le lien avec RSA ?

Exercice 7.14. Alice et Bob utilisent l'échange de clés de Diffie et Hellman avec $p = 23$ et $g = 5$.

a) Alice choisit $a = 6$ et Bob choisit $b = 15$. Quels nombres s'échangent-ils ? Quelle est leur clé commune ?

b) Ève a vu passer $A = 8$. Retrouver a en calculant les puissances successives de 5 modulo 23.

c) Pourquoi serait-il absurde de prendre $g = 1$? Et $g = 22$?

Exercice 7.15. On veut casser automatiquement un chiffre affine, sans intervention humaine.

a) Écrire une fonction `score(texte)` qui renvoie la somme des fréquences françaises (liste `FREQ` du programme 7.2) des lettres du texte.

b) Écrire une fonction `casser_affine(chiffre)` qui essaie les 312 clés de déchiffrement possibles et renvoie celle qui donne le texte de meilleur score, avec ce texte.

c) Appliquer la fonction au texte chiffré de l'exemple 7.6, puis aux messages `PCRZCQJ` et `MUZMCK` (exemple 7.5 et exercice 7.4). Commenter.

Exercice 7.16. On utilise le programme 7.2.

a) Choisir un texte français d'au moins 300 lettres (un article de journal, une page de roman) et le chiffrer par Vigenère avec un mot-clé de 4 à 8 lettres.

b) Échanger le texte chiffré avec une ou un collègue, sans lui communiquer la clé. Tracer, avec `matplotlib`, le diagramme à bandes du nombre de coïncidences pour les décalages de 1 à 20, et en déduire la longueur de la clé.

c) Retrouver la clé avec la fonction `trouver_cle`, puis déchiffrer.

d) Recommencer avec des textes de plus en plus courts. À partir de quelle longueur la méthode échoue-t-elle ?

Exercice 7.17. On programme une Enigma simplifiée, formée d'un seul rotor et d'un réflecteur, sans tableau de connexions. Le rotor est donné par son câblage, et le réflecteur par les lettres qu'il associe à $a, b, \ldots, z$:

```
ROTOR      = "ekmflgdqvzntowyhxuspaibrcj"
REFLECTEUR = "yruhqsldpxngokmiebfzcwvjat"
```

Le rotor envoie a sur e, b sur k, etc. ; c'est le câblage du rotor I des Enigma militaires.

a) Vérifier que le réflecteur échange les lettres deux à deux, sans point fixe.

b) Lorsque le rotor a tourné de k crans, il envoie la lettre x sur $R(x + k) - k$, où R est son câblage. Écrire une fonction `mini_enigma(texte, depart)` qui chiffre un texte, le rotor tournant d'un cran *avant* chaque lettre, à partir de la position `depart`.

c) Chiffrer `aaaaaaaaaaaaaaaaaaaaaaaaaa` (26 fois la lettre a). Qu'observe-t-on ?

d) Vérifier sur des exemples les deux propriétés de la proposition 7.3.

Exercice 7.18. Bob a publié la clé RSA $n = 1\,500\,011\,500\,021$ et $e = 65\,537$. Ève intercepte le message chiffré $c = 213\,152\,470\,940$. Le texte clair a été converti en un nombre en remplaçant chaque lettre par son numéro à deux chiffres ($a = 00$, $b = 01$, …, $z = 25$).

a) Écrire une fonction qui renvoie le plus petit diviseur premier d'un entier $n \geq 2$, par divisions successives. L'utiliser pour factoriser n.

b) En déduire la clé privée d, puis le message.

c) Combien de divisions la fonction a-t-elle effectuées ? En supposant qu'un ordinateur effectue un milliard de divisions par seconde, estimer le temps nécessaire pour factoriser ainsi un nombre $n = pq$ de 600 chiffres, où p et q ont 300 chiffres chacun.

Exercice 7.19. La cryptographie suscite facilement l'intérêt des élèves, et elle mobilise de nombreuses notions du programme du secondaire.

a) Relever quatre notions du programme du secondaire [Min16] qui interviennent dans ce chapitre, et préciser pour chacune le chiffrement ou l'attaque où elle intervient.

b) Concevoir les grandes lignes d'une activité de deux périodes, pour le premier cycle du secondaire, qui mène les élèves du chiffre de César à l'analyse des fréquences. Préciser l'intention pédagogique, le déroulement et ce que les élèves doivent produire.

c) Un élève affirme : « Si personne ne connaît ma méthode de chiffrement, mon message est en sécurité. » Rédiger une réponse de quelques lignes, adaptée à un élève du secondaire, à l'aide du principe de Kerckhoffs et d'un exemple tiré de l'histoire.

A Calculer les transformations avec Python

Cette annexe présente des programmes Python pour effectuer les calculs du chapitre 1. Ils permettent notamment de résoudre les deux systèmes 3×3 du problème des deux triangles. Nous utiliserons trois bibliothèques : `numpy` pour le calcul numérique [HMW+20] [1], `sympy` pour le calcul exact [MSP+17] [2] et `matplotlib` pour les figures [Hun07]. La lecture du chapitre ne demande pas d'utiliser ces programmes, mais ceux-ci permettent de reprendre les exemples et d'en explorer d'autres.

Représenter une figure et une transformation

Nous représenterons une figure par un tableau $2 \times n$ dont chaque *colonne* contient les coordonnées d'un point, suivant une convention courante en infographie. Cette disposition permet d'appliquer la partie linéaire d'une transformation à tous les points par un seul produit matriciel.

Programme A.1 : Appliquer une transformation affine à un nuage de points.

```
import numpy as np

def applique(A, t, pts):
    """A : matrice 2x2, t : vecteur (2,), pts : tableau 2 x n."""
    A = np.asarray(A, dtype=float)
    t = np.asarray(t, dtype=float).reshape(2, 1)
    return A @ np.asarray(pts, dtype=float) + t

# le rectangle [0,2] x [0,1], un sommet par colonne
rect = np.array([[0., 2., 2., 0.],
                 [0., 0., 1., 1.]])

A, t = [[1, -1], [2, 0]], [-1, 1]
print(applique(A, t, rect))
print("det =", np.linalg.det(A))
```

Programme A.2 : Sortie du programme.

```
[[-1.  1.  0. -2.]
 [ 1.  5.  5.  1.]]
det = 2.0000000000000004
```

Le calcul donne le parallélogramme de l'exemple 1.5. Il faut cependant interpréter la valeur affichée pour le déterminant : 2,0000000000000004 plutôt que 2. Cet écart vient de la représentation des nombres en virgule flottante. Le calcul *numérique* fournit en général des valeurs approchées, même lorsque la réponse mathématique est un entier.

1. Disponible en ligne (libre accès) : https://www.nature.com/articles/s41586-020-2649-2.
2. Disponible en ligne (libre accès) : https://doi.org/10.7717/peerj-cs.103.

Nous verrons plus loin comment utiliser le calcul symbolique pour obtenir une réponse exacte.

Une petite bibliothèque de transformations

Programme A.3 : Les matrices du catalogue de la section 1.6.

```
def rotation(theta):        # theta en radians
    c, s = np.cos(theta), np.sin(theta)
    return np.array([[c, -s], [s, c]])

def reflexion(theta):       # axe faisant l'angle theta avec Ox
    c, s = np.cos(2*theta), np.sin(2*theta)
    return np.array([[c, s], [s, -c]])

def homothetie(c):    return np.array([[c, 0.], [0., c]])
def cisaillement(m):  return np.array([[1., m], [0., 1.]])
```

Composer : les coordonnées homogènes en pratique

Les coordonnées homogènes de la section 1.7 permettent de composer les transformations par des produits de matrices 3×3. Reprenons la rotation r et la translation t de l'exemple 1.8. Le programme suivant calcule les matrices $M_t M_r$ et $M_r M_t$, que nous avons déjà obtenues à la main.

Programme A.4 : Coordonnées homogènes : composer, c'est multiplier.

```
def homogene(A, t):
    """Matrice 3x3 de w(X) = A X + t."""
    M = np.eye(3)
    M[:2, :2] = A
    M[:2, 2]  = t
    return M

def applique_h(M, pts):
    pts = np.asarray(pts, dtype=float)
    P = np.vstack([pts, np.ones(pts.shape[1])])
    return (M @ P)[:2]

M_rot  = homogene(rotation(np.pi/2), [0, 0])
M_tran = homogene(np.eye(2), [3, 1])

print(M_tran @ M_rot)   # d'abord la rotation, puis la translation
print(M_rot @ M_tran)   # l'ordre inverse
```

Programme A.5 : Sortie : la composition n'est pas commutative.

```
[[ 0. -1.  3.]        [[ 0. -1. -1.]
 [ 1.  0.  1.]         [ 1.  0.  3.]
 [ 0.  0.  1.]]        [ 0.  0.  1.]]
```

On constate que les deux matrices ont des parties translation différentes. Le calcul illustre donc la remarque 1.1 et l'exemple 1.4. Pour en faire une activité d'apprentissage, on peut demander aux élèves de *prévoir* les deux résultats avant d'exécuter le programme, puis de comparer leur prévision aux matrices obtenues.

Résoudre le problème des deux triangles

Le programme suivant résout le système (1.5) pour les données de l'exemple 1.10.

Programme A.6 : L'unique application affine envoyant trois points sur trois points.

```
def affine_par_trois_points(P, Q):
    """P, Q : listes de trois couples (x, y). Renvoie (A, t)."""
    M = np.array([[P[0][0], P[0][1], 1.],
                  [P[1][0], P[1][1], 1.],
                  [P[2][0], P[2][1], 1.]])
    if abs(np.linalg.det(M)) < 1e-12:
        raise ValueError("les trois points de depart sont alignes")
    abe = np.linalg.solve(M, [q[0] for q in Q])   # a, b, e
    cdf = np.linalg.solve(M, [q[1] for q in Q])   # c, d, f
    A = np.array([[abe[0], abe[1]], [cdf[0], cdf[1]]])
    t = np.array([abe[2], cdf[2]])
    return A, t

P = [(-11, 3), (1, 12), (-1, 14)]
Q = [(90, -68), (69, -53), (83, -59)]
A, t = affine_par_trois_points(P, Q)
print("A =", A.round(6).tolist())
print("t =", t.round(6).tolist())
```

Programme A.7 : Sortie.

```
A = [[-4.0, 3.0], [2.0, -1.0]]
t = [37.0, -43.0]
```

Le test portant sur le déterminant correspond à l'hypothèse de non-alignement des trois points dans le théorème 1.15. Il permet donc de vérifier que les données satisfont les conditions d'application du résultat. Cet exemple illustre un intérêt de la programmation en mathématiques : la rédaction du code amène à expliciter les hypothèses et à décider comment traiter les cas où elles ne sont pas satisfaites.

Des réponses exactes avec `sympy`

Pour conserver une réponse exacte contenant des fractions ou des expressions comme $\sqrt{3}/2$ et $\pi/4$, nous pouvons utiliser `sympy`. Cette bibliothèque effectue des calculs symboliques, sans remplacer systématiquement les nombres par des approximations décimales.

Programme A.8 : Le même problème, en calcul exact.

```
import sympy as sp

a, b, c, d, e, f = sp.symbols('a b c d e f')
P = [(-11, 3), (1, 12), (-1, 14)]
Q = [(90, -68), (69, -53), (83, -59)]

eqs = []
for (x, y), (u, v) in zip(P, Q):
    eqs.append(sp.Eq(a*x + b*y + e, u))
    eqs.append(sp.Eq(c*x + d*y + f, v))

print(sp.solve(eqs, [a, b, c, d, e, f], dict=True)[0])
```

Programme A.9 : Sortie : des entiers, pas des approximations.

```
{a: -4, b: 3, c: 2, d: -1, e: 37, f: -43}
```

Le calcul symbolique permet aussi de vérifier des identités. Le code suivant établit la relation $A^\top A = k^2 I$ pour la matrice d'une similitude directe.

Programme A.10 : Vérification symbolique de $A^\top A = k^2 I$.

```
k, th = sp.symbols('k theta', positive=True)
S = k * sp.Matrix([[sp.cos(th), -sp.sin(th)],
                   [sp.sin(th),  sp.cos(th)]])
print(sp.simplify(S.T * S))     # Matrix([[k**2, 0], [0, k**2]])
print(sp.simplify(S.det()))     # k**2
```

Reconnaître une similitude

Le critère $A^\top A = k^2 I$ permet d'écrire un programme qui vérifie si une matrice représente une similitude.

Programme A.11 : Est-ce une similitude ? Si oui, de quel rapport et de quel genre ?

```
def est_similitude(A, tol=1e-9):
    A = np.asarray(A, dtype=float)
    G = A.T @ A                     # doit valoir k^2 * I
    k2 = G[0, 0]
    if k2 <= tol:
        return (False, None, None)
    if np.linalg.norm(G - k2*np.eye(2)) > tol * max(1.0, k2):
        return (False, None, None)
    genre = "directe" if np.linalg.det(A) > 0 else "indirecte"
    return (True, float(np.sqrt(k2)), genre)

print(est_similitude([[0, -0.5], [0.5, 0]]))   # rotation + homothetie
print(est_similitude([[0.8, 0], [0.8, 0.6]]))  # exemple 1.9
print(est_similitude([[-4, 3], [2, -1]]))      # les deux triangles
```

Programme A.12 : Sortie.

```
(True, 0.5, 'directe')
(False, None, None)
(False, None, None)
```

Le deuxième appel retrouve le résultat de l'exemple 1.9 : aucune valeur du paramètre ne fait de la matrice $\begin{pmatrix} 0{,}8 & 0 \\ a & 0{,}6 \end{pmatrix}$ une similitude.

Dessiner

Nous utiliserons enfin `matplotlib` pour représenter la figure de départ et son image.

Programme A.13 : Tracer une figure et son image.

```
import matplotlib.pyplot as plt

maison = np.array([[0., 2., 2., 1., 0.],
                   [0., 0., 1.4, 2.2, 1.4]])
```

```
image = applique([[1, -1], [2, 0]], [-1, 1], maison)

fig, ax = plt.subplots(figsize=(5.4, 3.4))
for pts, coul, lab in [(maison, "#36648B", "figure $F$"),
                       (image,  "#CD3333", "$w(F)$")]:
    ferme = np.hstack([pts, pts[:, :1]])       # on referme le polygone
    ax.fill(ferme[0], ferme[1], color=coul, alpha=0.25)
    ax.plot(ferme[0], ferme[1], color=coul, lw=1.8, label=lab)

ax.set_aspect("equal")          # indispensable : sinon la figure ment
ax.grid(alpha=0.25)
ax.legend(frameon=False)
plt.show()
```

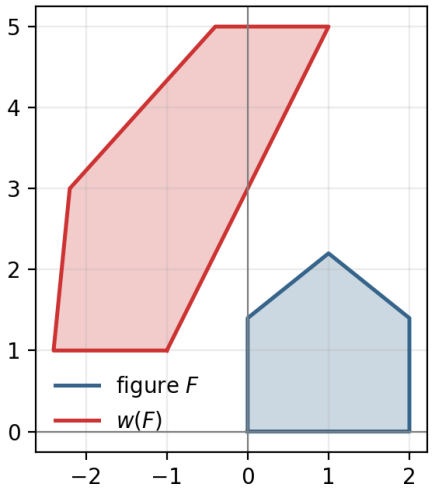


FIGURE A.1 : Figure produite par le programme précédent. Le motif en forme de maison est déformé : les longueurs et les angles ne sont pas conservés, mais le parallélisme et l'alignement le sont, conformément à la proposition 1.3.

Remarque A.1 **Le choix de l'échelle des axes**

La commande `ax.set_aspect("equal")` impose une même échelle sur les deux axes. Si on l'omet, `matplotlib` peut les ajuster séparément pour remplir la fenêtre. Un cercle apparaît alors comme une ellipse et un angle droit peut sembler ne plus l'être. On risque ainsi d'attribuer à la transformation une déformation qui vient uniquement de l'affichage. Cet exemple donne une occasion de discuter en classe du rôle des échelles et de la nécessité d'interpréter les réglages d'un outil graphique.

Reconnaître le genre et l'angle d'une similitude

Le programme suivant répond à la demande de l'exercice 1.17. Il s'appuie sur le fait qu'une matrice orthogonale 2×2 est nécessairement la matrice d'une rotation ou d'une réflexion : une fois vérifié que $A^\top A = k^2 I$, on pose $Q = A/k$; si $\det(Q) > 0$, Q est une rotation d'angle $\theta = \operatorname{atan2}(Q_{21}, Q_{11})$, et sinon une réflexion dont l'axe fait l'angle $\frac{1}{2}\operatorname{atan2}(Q_{21}, Q_{11})$ avec Ox, d'après l'équation (1.2).

Programme A.14 : Décomposer une similitude en rapport, angle et genre.

```
import numpy as np
```

```
def decompose(A):
    A = np.asarray(A, dtype=float)
    k2 = (A.T @ A)[0, 0]
    if not np.allclose(A.T @ A, k2*np.eye(2)):
        raise ValueError("ce n'est pas une similitude")
    k = np.sqrt(k2)
    Q = A / k
    if np.linalg.det(Q) > 0:
        return k, np.arctan2(Q[1, 0], Q[0, 0]), "directe"
    return k, 0.5*np.arctan2(Q[1, 0], Q[0, 0]), "indirecte"
```

On peut vérifier le fonctionnement du programme avec les matrices de la table 1.1, puis comparer les résultats aux propriétés géométriques indiquées dans cette table.

B La rédaction du chapitre 2 avec l'assistance d'une intelligence artificielle

Cette annexe décrit la démarche suivie pour rédiger le chapitre 2, à partir d'échanges avec Claude, un assistant conversationnel développé par Anthropic [1]. L'auteur a fourni les documents de départ, précisé les attentes, relu les versions et demandé les corrections. L'assistant a produit le texte et les figures, puis appliqué ces demandes. La responsabilité du contenu mathématique et des choix pédagogiques appartient à l'auteur. Les extraits qui suivent reproduisent les demandes ayant guidé ce travail, avec quelques corrections de frappe et de langue.

Les documents de départ

Deux documents ont servi de base à la rédaction. Le fichier `prompt.pdf`, préparé par l'auteur, précisait les notions à traiter et leur importance relative. Il contenait aussi une démonstration manuscrite du théorème de Banach à reprendre dans le chapitre. Le fichier `EspacesMetriquesan.pdf` était une version antérieure des notes de l'auteur sur les espaces métriques. Plusieurs définitions, exemples et figures en ont été tirés.

Le mandat initial

La demande initiale précisait les objectifs et les grandes lignes du chapitre.

> Je veux que tu rédiges le chapitre 2 sur les espaces métriques, pour en arriver à présenter la distance de Hausdorff et le théorème de Banach. Il faudra présenter un peu plus de métriques appliquées ; tu verras que j'ai mentionné une métrique sphérique et une autre pour mesurer les erreurs dans les chaînes binaires. Commence par comprendre le PDF `prompt.pdf` et pose-moi des questions pour valider ta compréhension. J'ai aussi une ébauche de notes sur les espaces métriques (`EspacesMetriquesan.pdf`) que tu peux regarder pour te donner une idée. Ça doit être très bien illustré et vulgarisé, pour des étudiantes et étudiants du baccalauréat en enseignement. En termes de répartition : un tiers sur la notion d'espace métrique, avec exemples ; un tiers sur tout ce qui touche à Hausdorff et Minkowski ; un tiers sur la notion de point fixe et le théorème du point fixe de Banach. Je t'ai fourni la démonstration que je juge la plus appropriée pour ce cours ; j'attends ton avis à ce sujet. On s'attend probablement à une trentaine de pages pour ce chapitre. N'oublie pas la biographie de Félix Hausdorff et, tant qu'à y être, une sur Banach. Ce serait bien de parler de l'historique de ce théorème, qui a des applications dans plusieurs domaines.

1. Plus précisément, le modèle Claude Sonnet 5.

Valider la compréhension avant d'écrire

Après avoir lu les deux documents, l'assistant a posé des questions sur l'organisation du travail : emplacement du chapitre dans le fichier source, illustration des biographies et contenu des exemples de métriques appliquées. D'autres questions ont suivi la rédaction d'une première version et sa relecture. Les deux réponses ci-dessous, formulées librement plutôt que choisies parmi des options, montrent le type de précisions apportées.

> En gros, je veux que tu évites le tiret cadratin dans une phrase où il y a des maths (en mode mathématique). Tu peux mettre des virgules, des deux-points, ou un autre arrangement.

> Je veux qu'on traite la métrique sphérique de manière qualitative seulement : des explications du fait qu'on ne peut pas traverser la Terre, qu'on doit se déplacer à sa surface, avec une représentation graphique (en s'assurant que l'arc de grand cercle est bien au-dessus de la corde sur le graphique).

Une relecture détaillée

La première version complète a ensuite été relue section par section. Les commentaires portaient à la fois sur des corrections locales et sur l'organisation du contenu. Les extraits suivants en donnent quelques exemples.

> Le texte en bleu dans la figure devrait être en haut, car il chevauche les courbes.

> Ne fais pas la définition formelle, mais présente seulement la distance sphérique avec la figure (partie de gauche) ; la partie de droite peut être retirée. Enlève l'exemple numérique : on le gardera, tout comme la définition rigoureuse, pour le chapitre sur le positionnement, où les coordonnées sphériques seront présentées.

> Pour les boules ouvertes et les ensembles ouverts en général, la frontière doit être en pointillé.

> Pour l'exemple sur la complétude de $\mathbb{R}$, il faut construire une suite convergente sans utiliser la méthode de Newton-Raphson, puisqu'elle n'est pas encore vue à ce stade. Un raisonnement basé sur les décimales de la racine de 2 suffira peut-être.

> Illustrons, pour les fonctions réelles d'une variable réelle, graphiquement ce que signifie être un point fixe.

> Il faut réécrire le théorème de Banach dans le cas particulier où l'espace est $\mathbb{R}^2$ avec la distance euclidienne, mais en ajoutant une remarque précisant que la démonstration reste la même si on remplace $\mathbb{R}^2$ par un espace métrique quelconque. On procède ainsi par souci pédagogique, pour pouvoir introduire des images, notamment pour l'application répétée de l'inégalité triangulaire et pour le lemme selon lequel une contraction de rapport k, itérée n fois, en est une de rapport k^n.

> Pour la méthode de Newton-Raphson, en plus de l'exemple théorique, ajoute un problème plus concret : un problème d'optimisation qui mène à une équation polynomiale de degré 3 dont les racines ne sont pas évidentes à trouver. Je pense à un problème d'optimisation des dimensions d'un solide, avec une condition d'aire minimale, ou à une clôture avec contrainte de périmètre.

Raffiner les figures

Les figures ont aussi fait l'objet de plusieurs échanges. Une capture d'écran annotée permettait souvent de désigner précisément l'élément à modifier.

> Pour la figure du ε-voisinage, mets une double flèche avec ε en dessous, à peu près où j'ai mis l'annotation : ce sera plus clair que ta flèche à sens unique.

> Je pense que ça devrait être un peu plus gros comme graphique ; les étiquettes ne devraient pas se superposer aux courbes, et peut-être que l'utilisation du package `subfigure` de LaTeX serait appropriée pour gérer les trois volets (i), (ii) et (iii).

Restructurer le contenu

Les deux demandes suivantes ont conduit à revoir plus largement l'organisation du chapitre.

> Ce serait mieux d'utiliser un carré pour représenter $\mathbb{R}^2$ qu'une forme irrégulière.

> Toutes les discussions entourant les fractales doivent être déplacées au chapitre 3. Tu peux simplement mentionner que les fractales seront définies comme des objets dans l'ensemble des compacts du plan, ce qui motive cette construction.

Les détails techniques

D'autres échanges concernaient la production du document, notamment la compilation et les conventions typographiques.

> Peux-tu t'assurer, dans la configuration de VS Code et dans ta propre façon de faire, de toujours compiler avec LuaLaTeX et avec SyncTeX ? Ce n'est pas le cas actuellement.

> Assure-toi que les blocs de définition, lemme, corollaire et théorème soient non sécables (*unbreakable*), car présentement un des lemmes déborde d'une page à l'autre.

> Je veux utiliser la commande `\num` du package `siunitx`, car en français, c'est la virgule qui doit servir de séparateur décimal.

En rétrospective

Cette démarche a permis de relire et de corriger les sections au fur et à mesure de leur rédaction. Elle a aussi demandé une vérification suivie de la part de l'auteur. La rapidité de production de l'assistant ne dispense pas d'examiner les raisonnements, les figures et la pertinence des explications pour les étudiants. Les décisions mathématiques et pédagogiques demeurent donc sous la responsabilité de l'auteur.

C Compléments théoriques sur les SIF

Cette annexe rassemble les démonstrations du chapitre 3 présentées séparément afin d'alléger le texte principal. Nous montrerons d'abord que l'opérateur de Hutchinson d'un SIF est une contraction de l'espace $(\mathcal{K}(\mathbb{R}^2), d_H)$ (théorème 3.1), puis nous en déduirons le théorème de Hutchinson (théorème 3.3) à l'aide du théorème de Banach. Les démonstrations ne font appel qu'aux ε-voisinages de la définition 2.8. Rappelons que $d_H(A,B) \leq \varepsilon$ signifie exactement que $A \subseteq B^{\oplus\varepsilon}$ et $B \subseteq A^{\oplus\varepsilon}$ (l'infimum de la définition 2.9 étant atteint).

Lemme C.1 **Distance entre deux réunions**

Soient A, B, C et D des compacts non vides du plan. Alors

$$d_H(A \cup B,\ C \cup D) \leq \max\{d_H(A,C),\ d_H(B,D)\}.$$

Démonstration. Posons $\varepsilon = \max\{d_H(A,C), d_H(B,D)\}$. Puisque $d_H(A,C) \leq \varepsilon$, on a $A \subseteq C^{\oplus\varepsilon}$; de même, $B \subseteq D^{\oplus\varepsilon}$. Or un point à distance au plus ε de C, ou de D, est à distance au plus ε de $C \cup D$: $C^{\oplus\varepsilon}$ et $D^{\oplus\varepsilon}$ sont tous deux contenus dans $(C \cup D)^{\oplus\varepsilon}$. Ainsi $A \cup B \subseteq (C \cup D)^{\oplus\varepsilon}$. En reprenant le raisonnement avec les rôles inversés, nous obtenons $C \cup D \subseteq (A \cup B)^{\oplus\varepsilon}$, d'où $d_H(A \cup B, C \cup D) \leq \varepsilon$. □

Lemme C.2 **Une contraction rapproche les ensembles**

Soit w une contraction du plan, de rapport k, et soient A et B deux compacts non vides. Alors

$$d_H(w(A), w(B)) \leq k\, d_H(A,B).$$

Démonstration. Posons $\varepsilon = d_H(A,B)$ et prenons un point quelconque de $w(A)$; il s'écrit $w(a)$ pour un certain $a \in A$. Puisque $A \subseteq B^{\oplus\varepsilon}$, il existe $b \in B$ tel que $d_E(a,b) \leq \varepsilon$. Alors $w(b) \in w(B)$ et

$$d_E(w(a), w(b)) \leq k\, d_E(a,b) \leq k\varepsilon,$$

de sorte que $w(a) \in w(B)^{\oplus k\varepsilon}$. Nous avons donc établi que $w(A) \subseteq w(B)^{\oplus k\varepsilon}$; en échangeant les rôles de A et B, on obtient $w(B) \subseteq w(A)^{\oplus k\varepsilon}$, d'où le résultat. □

Démonstration du théorème 3.1

Démonstration. Le lemme C.1 s'étend par récurrence à une réunion de N ensembles. On a donc

$$\begin{aligned} d_H(W(A), W(B)) &= d_H\Big(\bigcup_{i=1}^{N} w_i(A),\ \bigcup_{i=1}^{N} w_i(B)\Big) \leq \max_{1\leq i\leq N} d_H(w_i(A), w_i(B)) \\ &\leq \max_{1\leq i\leq N} k_i\, d_H(A,B), \end{aligned}$$

La dernière inégalité découle du lemme C.2. Le membre de droite vaut $k\, d_H(A, B)$. □

Démonstration du théorème 3.3

Démonstration. Par les théorèmes 3.1 et 3.2, W est une contraction de rapport k de l'espace métrique complet $\left(\mathcal{K}(\mathbb{R}^2), d_H\right)$. Le théorème de Banach, sous la forme générale de la remarque 2.11, s'applique donc à $f = W$: les points (i), (ii) et (iii) correspondent à ses conclusions exprimées dans cet espace, avec $x_0 = B$ et $x_1 = W(B)$. □

D Rappels d'algèbre linéaire

Cette annexe rassemble les notions d'algèbre linéaire sur lesquelles s'appuient les chapitres du livre. Il ne s'agit pas d'un cours complet : la plupart des démonstrations sont omises, et nous renvoyons à un manuel d'algèbre linéaire pour les détails. L'objectif est plutôt de rafraîchir la mémoire, de fixer les notations et, surtout, de redonner une *image géométrique* à des calculs qui ont souvent été appris comme des recettes.

L'ordre de présentation suit la façon dont les notions s'appuient les unes sur les autres. Nous partirons des vecteurs, que nous apprendrons à combiner et à comparer à l'aide du produit scalaire. Nous rangerons ensuite des vecteurs dans des matrices, ce qui nous permettra de résoudre des systèmes linéaires par la méthode de Gauss–Jordan. Les systèmes nous mèneront à l'indépendance linéaire, aux bases et au noyau. Nous verrons ensuite comment le déterminant mesure l'effet d'une matrice sur les aires et les volumes, et quand une matrice est inversible. Nous terminerons par les valeurs propres et la diagonalisation, qui répondent à une question très géométrique : quelles directions une matrice laisse-t-elle inchangées ?

Les exemples sont volontairement courts et presque toujours en dimension 2 ou 3, où l'on peut faire un dessin. Dès que les calculs deviennent longs, nous les confierons à Python, avec les bibliothèques `numpy` et `sympy` déjà présentées à l'annexe A.

D.1 Vecteurs de $\mathbb{R}^n$

Définition D.1 **Vecteur, opérations**

Un **vecteur** de $\mathbb{R}^n$ est une colonne de n nombres réels, appelés ses **composantes** :

$$\mathbf{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}, \qquad \text{que l'on écrit aussi } \mathbf{u} = (u_1, \dots, u_n).$$

L'**addition** de deux vecteurs et la **multiplication par un scalaire** $c \in \mathbb{R}$ se font composante par composante :

$$\mathbf{u} + \mathbf{v} = (u_1 + v_1, \dots, u_n + v_n), \qquad c\,\mathbf{u} = (cu_1, \dots, cu_n).$$

Le **vecteur nul** est $\mathbf{0} = (0, \dots, 0)$.

Ces deux opérations obéissent aux règles de calcul habituelles : l'addition est commutative et associative, $\mathbf{u} + \mathbf{0} = \mathbf{u}$, et l'on a $c(\mathbf{u} + \mathbf{v}) = c\mathbf{u} + c\mathbf{v}$ ainsi que $(c + d)\mathbf{u} = c\mathbf{u} + d\mathbf{u}$. Un ensemble muni de deux opérations qui respectent ces règles s'appelle un *espace vectoriel*. Les polynômes, les fonctions continues ou les images numériques (vues comme des

tableaux de nombres) en sont d'autres exemples, mais $\mathbb{R}^n$ nous suffira ici.

Géométriquement, comme à la remarque 1.4, on représente $\mathbf{u} \in \mathbb{R}^2$ ou $\mathbb{R}^3$ par une flèche partant de l'origine, ou par le point situé au bout de cette flèche. L'addition correspond alors à la *règle du parallélogramme* (ou, ce qui revient au même, à la mise bout à bout des flèches), et la multiplication par c allonge ou raccourcit la flèche d'un facteur $|c|$, en la retournant si $c < 0$ (figure D.1).

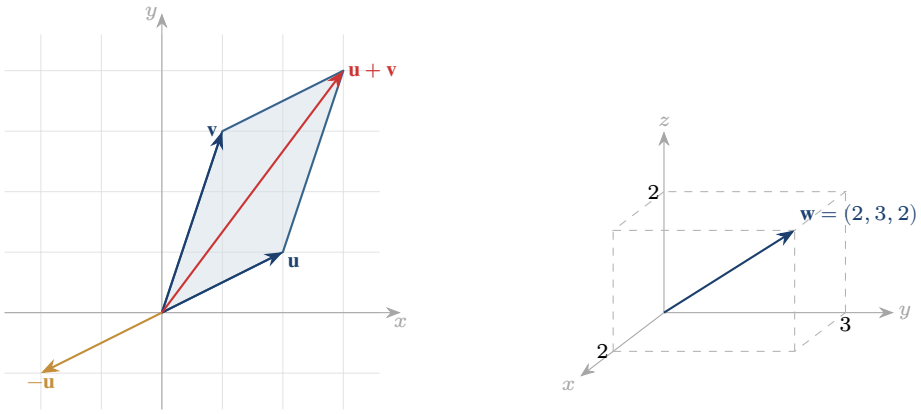


FIGURE D.1 : À gauche, dans $\mathbb{R}^2$: la somme de $\mathbf{u} = (2, 1)$ et de $\mathbf{v} = (1, 3)$ est la diagonale du parallélogramme qu'ils engendrent ; $-\mathbf{u}$ a la même longueur que $\mathbf{u}$, mais pointe dans le sens opposé. À droite, dans $\mathbb{R}^3$: les trois composantes d'un vecteur se lisent sur les arêtes d'une boîte.

Pour la classe D.1 — **Les vecteurs au secondaire**

En cinquième secondaire (séquence SN), les élèves manipulent déjà les vecteurs du plan : somme par la règle du parallélogramme ou de Chasles, multiplication par un scalaire, combinaisons linéaires et produit scalaire. Ils les notent souvent $\overrightarrow{AB}$ et les représentent d'abord par des flèches, avant de passer aux composantes. Tout ce qui suit prolonge donc des notions familières ; la nouveauté tient surtout au passage à $\mathbb{R}^n$, où l'on ne peut plus dessiner, mais où les calculs restent les mêmes.

D.2 Combinaisons linéaires et espace engendré

Définition D.2 — **Combinaison linéaire, espace engendré**

Une **combinaison linéaire** des vecteurs $\mathbf{v}_1, \dots, \mathbf{v}_k$ est un vecteur de la forme

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_k\mathbf{v}_k, \qquad c_1, \dots, c_k \in \mathbb{R}\,.$$

L'ensemble de toutes ces combinaisons s'appelle l'**espace engendré** par $\mathbf{v}_1, \dots, \mathbf{v}_k$; on le note $\operatorname{span}\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$.

Remarque D.1 **Span ou Vect ?**

La notation span n'est pas un acronyme : c'est le mot anglais *span*, qui signifie « étendue » ou « envergure ». L'espace engendré, c'est tout ce que les vecteurs permettent d'atteindre. Elle est d'usage courant au Québec, même dans les textes rédigés en français. En France, on écrit plutôt $\mathrm{Vect}\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$, abréviation de « sous-espace *vect*oriel engendré ». Certains manuels utilisent aussi $\langle \mathbf{v}_1, \dots, \mathbf{v}_k \rangle$. Toutes ces notations désignent le même ensemble.

L'espace engendré par un seul vecteur non nul est la droite qui passe par l'origine dans sa direction. L'espace engendré par deux vecteurs non parallèles de $\mathbb{R}^3$ est un plan passant par l'origine : on s'y déplace en avançant « un peu selon $\mathbf{u}$ et un peu selon $\mathbf{v}$ » (figure D.2). Dans tous les cas, un espace engendré contient $\mathbf{0}$ (on prend tous les c_i nuls) et reste stable par addition et par multiplication par un scalaire ; on dit que c'est un **sous-espace vectoriel**.

Exemple D.1 **Est-ce une combinaison linéaire ?**

Dans $\mathbb{R}^2$, le vecteur $\mathbf{w} = (4, 7)$ est-il une combinaison linéaire de $\mathbf{u} = (2, 1)$ et de $\mathbf{v} = (1, 3)$? On cherche a et b tels que $a\mathbf{u} + b\mathbf{v} = \mathbf{w}$, c'est-à-dire

$$\begin{cases} 2a + b = 4 \\ a + 3b = 7. \end{cases}$$

La première équation donne $b = 4 - 2a$; en substituant, $a + 12 - 6a = 7$, d'où $a = 1$ et $b = 2$. Ainsi, $\mathbf{w} = \mathbf{u} + 2\mathbf{v}$.
Dans $\mathbb{R}^3$, prenons maintenant $\mathbf{u} = (1, 0, 1)$ et $\mathbf{v} = (0, 1, 1)$. Une combinaison $a\mathbf{u} + b\mathbf{v} = (a, b, a + b)$ a toujours sa troisième composante égale à la somme des deux premières. Le vecteur $(1, 1, 2)$ est donc dans $\mathrm{span}\{\mathbf{u}, \mathbf{v}\}$, mais pas $(1, 1, 3)$: ce plan est celui d'équation $z = x + y$.

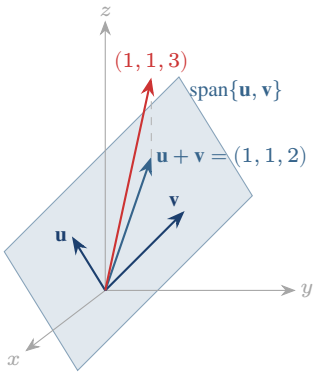


FIGURE D.2 : Les combinaisons linéaires de $\mathbf{u} = (1, 0, 1)$ et de $\mathbf{v} = (0, 1, 1)$ remplissent le plan $z = x + y$. Le vecteur $(1, 1, 3)$ sort de ce plan : ce n'est pas une combinaison linéaire de $\mathbf{u}$ et de $\mathbf{v}$.

Il faut bien voir que la question de l'exemple D.1 s'est ramenée à un *système d'équations linéaires* dont les inconnues sont les coefficients. C'est le fil conducteur de toute l'annexe : presque toutes les questions sur les vecteurs finissent par se traduire en systèmes, que nous apprendrons à résoudre efficacement à la section D.5.

D.3 Produit scalaire, norme et angle

Définition D.3 **Produit scalaire, norme**

Le **produit scalaire** de $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ est le nombre

$$\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + u_2 v_2 + \cdots + u_n v_n.$$

La **norme** (ou longueur) de $\mathbf{u}$ est $\|\mathbf{u}\| = \sqrt{\mathbf{u} \cdot \mathbf{u}}$. Deux vecteurs sont **orthogonaux** si $\mathbf{u} \cdot \mathbf{v} = 0$, et un vecteur de norme 1 est dit **unitaire**.

Le produit scalaire est symétrique ($\mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u}$), linéaire en chacun de ses arguments, et $\mathbf{u} \cdot \mathbf{u} > 0$ dès que $\mathbf{u} \neq \mathbf{0}$. Dans $\mathbb{R}^2$ et $\mathbb{R}^3$, la norme n'est rien d'autre que la longueur de la flèche, calculée par Pythagore ; c'est la distance euclidienne d_E du chapitre 2. Pour rendre unitaire un vecteur non nul, il suffit de le diviser par sa norme : $\mathbf{u}/\|\mathbf{u}\|$.

L'intérêt géométrique du produit scalaire vient de la formule

$$\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\| \, \|\mathbf{v}\| \cos\theta, \tag{D.1}$$

où $\theta \in [0, \pi]$ est l'angle entre les deux vecteurs. Elle découle de la loi des cosinus appliquée au triangle de côtés $\mathbf{u}$, $\mathbf{v}$ et $\mathbf{u} - \mathbf{v}$. Puisque $|\cos\theta| \leq 1$, on en tire l'**inégalité de Cauchy–Schwarz**, $|\mathbf{u} \cdot \mathbf{v}| \leq \|\mathbf{u}\| \, \|\mathbf{v}\|$, qui reste vraie dans $\mathbb{R}^n$ et sert justement à *définir* l'angle quand on ne peut plus le dessiner. Le signe du produit scalaire indique d'un coup d'œil si l'angle est aigu (> 0), droit ($= 0$) ou obtus (< 0).

Définition D.4 **Projection orthogonale**

La **projection orthogonale** de $\mathbf{v}$ sur la droite engendrée par un vecteur $\mathbf{u} \neq \mathbf{0}$ est

$$\operatorname{proj}_{\mathbf{u}} \mathbf{v} = \frac{\mathbf{v} \cdot \mathbf{u}}{\mathbf{u} \cdot \mathbf{u}} \, \mathbf{u}.$$

Le vecteur $\mathbf{v} - \operatorname{proj}_{\mathbf{u}} \mathbf{v}$ est orthogonal à $\mathbf{u}$.

Exemple D.2 **Un angle et une projection**

Reprenons $\mathbf{u} = (2, 1)$ et $\mathbf{v} = (1, 3)$. On a $\mathbf{u} \cdot \mathbf{v} = 2 + 3 = 5$, $\|\mathbf{u}\| = \sqrt{5}$ et $\|\mathbf{v}\| = \sqrt{10}$, d'où

$$\cos\theta = \frac{5}{\sqrt{5}\,\sqrt{10}} = \frac{1}{\sqrt{2}}, \qquad \theta = \frac{\pi}{4}.$$

La projection de $\mathbf{v}$ sur la droite de $\mathbf{u}$ vaut $\frac{5}{5}\,\mathbf{u} = (2, 1)$: ici, le hasard fait bien les choses, et le pied de la perpendiculaire tombe exactement au bout de $\mathbf{u}$. On vérifie que $\mathbf{v} - \mathbf{u} = (-1, 2)$ est bien orthogonal à $\mathbf{u}$: $(-1)(2) + (2)(1) = 0$ (figure D.3).

La projection orthogonale apparaît partout : pour calculer la distance d'un point à une droite, pour décomposer une force en composantes, ou encore, en statistique, pour ajuster une droite de régression par la méthode des moindres carrés, qui revient à projeter un vecteur de données sur un sous-espace.

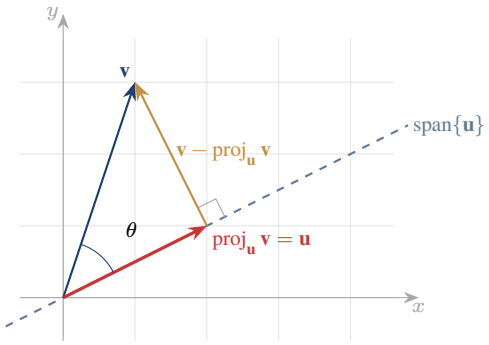


FIGURE D.3 : La projection orthogonale de $\mathbf{v}$ sur la droite engendrée par $\mathbf{u}$. Le vecteur « résidu » $\mathbf{v} - \text{proj}_{\mathbf{u}}\,\mathbf{v}$ est perpendiculaire à cette droite : la projection est le point de la droite le plus proche de $\mathbf{v}$.

D.4 Matrices et opérations matricielles

Une **matrice** $m \times n$ est un tableau de nombres à m lignes et n colonnes. On note a_{ij} le coefficient situé à la ligne i et à la colonne j, et $A = (a_{ij})$. Une matrice $n \times n$ est dite **carrée** ; la **matrice identité** I_n est la matrice carrée qui a des 1 sur la diagonale et des 0 ailleurs. On peut voir une matrice comme une rangée de n vecteurs colonnes de $\mathbb{R}^m$, et c'est souvent le point de vue le plus éclairant :

$$A = \left(\mathbf{a}_1 \;\middle|\; \mathbf{a}_2 \;\middle|\; \cdots \;\middle|\; \mathbf{a}_n \right).$$

L'addition de deux matrices de même taille et la multiplication par un scalaire se font coefficient par coefficient, exactement comme pour les vecteurs. La **transposée** $A^\intercal$ est la matrice $n \times m$ obtenue en échangeant les lignes et les colonnes : $(A^\intercal)_{ij} = a_{ji}$. Une matrice carrée est **symétrique** si $A^\intercal = A$.

D.4.1 Le produit d'une matrice par un vecteur

Si A est $m \times n$ et $\mathbf{x} \in \mathbb{R}^n$, on définit

$$A\mathbf{x} = x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n \in \mathbb{R}^m\,. \tag{D.2}$$

Autrement dit, $A\mathbf{x}$ *est la combinaison linéaire des colonnes de* A *dont les coefficients sont les composantes de* $\mathbf{x}$. En pratique, on calcule la i^{e} composante de $A\mathbf{x}$ en faisant le produit scalaire de la i^{e} ligne de A avec $\mathbf{x}$; les deux lectures donnent évidemment le même résultat. Par exemple,

$$\begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = 1 \begin{pmatrix} 2 \\ 1 \end{pmatrix} + 2 \begin{pmatrix} 1 \\ 3 \end{pmatrix} = \begin{pmatrix} 4 \\ 7 \end{pmatrix},$$

ce qui n'est rien d'autre que le calcul de l'exemple D.1, écrit autrement.

L'application $T(\mathbf{x}) = A\mathbf{x}$ est **linéaire** : $T(\mathbf{x} + \mathbf{y}) = T(\mathbf{x}) + T(\mathbf{y})$ et $T(c\mathbf{x}) = cT(\mathbf{x})$. Réciproquement, toute application linéaire de $\mathbb{R}^n$ dans $\mathbb{R}^m$ s'écrit ainsi, et ses colonnes sont les images des vecteurs de la base standard, $\mathbf{a}_j = T(\mathbf{e}_j)$. C'est exactement ce que montrait la figure 1.4 du premier chapitre : pour comprendre ce que fait une matrice 2×2, il suffit de regarder où elle envoie le carré unité.

D.4.2 Le produit de deux matrices

Définition D.5 **Produit matriciel**

Si A est $m \times n$ et B est $n \times p$, le produit AB est la matrice $m \times p$ dont le coefficient (i, j) est le produit scalaire de la ligne i de A par la colonne j de B :

$$(AB)_{ij} = \sum_{k=1}^{n} a_{ik}\, b_{kj}.$$

De façon équivalente, la colonne j de AB est $A\mathbf{b}_j$.

Cette définition, qui paraît arbitraire la première fois qu'on la rencontre, est taillée sur mesure pour la *composition* : $(AB)\mathbf{x} = A(B\mathbf{x})$. Appliquer AB, c'est d'abord appliquer B, puis A. On retrouve ainsi une idée centrale du chapitre 1 : composer des transformations linéaires, c'est multiplier leurs matrices (et c'est pour garder cette règle avec les translations qu'on passe aux coordonnées homogènes de la section 1.7).

Exemple D.3 **L'ordre compte**

Soit $C = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$ (un cisaillement) et $R = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ (la rotation d'un quart de tour). Alors

$$CR = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 2 & -1 \\ 1 & 0 \end{pmatrix},$$
$$RC = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 2 \end{pmatrix}.$$

Tourner puis cisailler ne donne pas le même résultat que cisailler puis tourner : le produit matriciel n'est pas commutatif.

Retenons les règles de calcul suivantes, valables dès que les tailles sont compatibles : le produit est associatif, $(AB)C = A(BC)$; il est distributif, $A(B + C) = AB + AC$; $I_m A = A I_n = A$; et $(AB)^\intercal = B^\intercal A^\intercal$, avec un changement d'ordre à ne pas oublier. En revanche, il faut se méfier de deux réflexes hérités des nombres : en général $AB \neq BA$, et $AB = 0$ n'entraîne pas que $A = 0$ ou $B = 0$ (essayez avec $A = B = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$).

D.5 Systèmes linéaires et méthode de Gauss–Jordan

Un **système linéaire** de m équations à n inconnues s'écrit

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ \qquad\qquad\qquad\qquad\quad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m, \end{cases} \quad \text{ou, en abrégé,} \quad A\mathbf{x} = \mathbf{b}.$$

D'après (D.2), résoudre ce système revient à chercher comment écrire $\mathbf{b}$ comme combinaison linéaire des colonnes de A. On peut aussi le lire ligne par ligne : chaque équation décrit une droite (si $n = 2$) ou un plan (si $n = 3$), et les solutions sont les points communs à toutes ces figures.

D.5.1 Ce que la géométrie nous apprend

Deux droites du plan se coupent en un point, sont parallèles et distinctes, ou sont confondues. Un système de deux équations à deux inconnues a donc une solution unique, aucune solution, ou une infinité de solutions (figure D.4). En dimension 3, les configurations de plans sont plus variées (figure D.5), mais la conclusion est la même, et elle reste vraie en toute dimension : *un système linéaire a zéro, une ou une infinité de solutions*, jamais exactement deux ou trois.

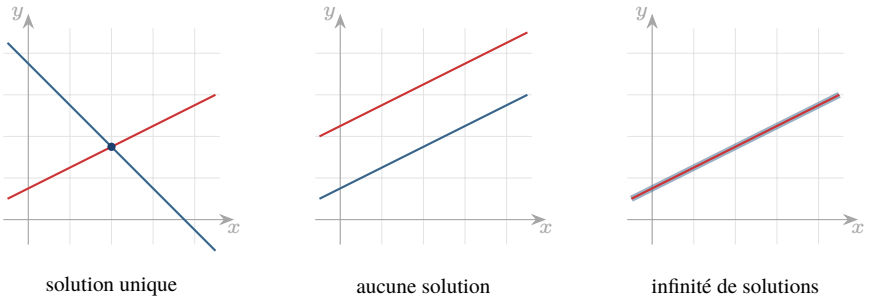


FIGURE D.4 : Les trois comportements possibles d'un système de deux équations à deux inconnues : droites sécantes, parallèles distinctes ou confondues.

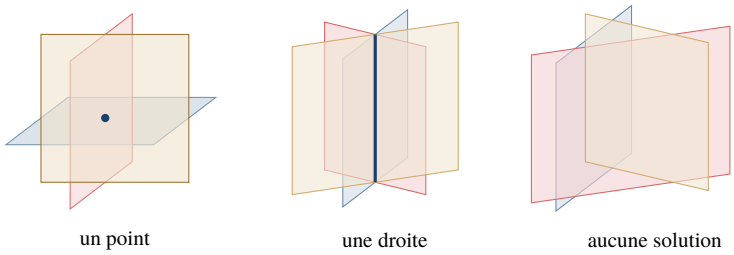


FIGURE D.5 : Trois équations à trois inconnues : trois plans de $\mathbb{R}^3$. Ils peuvent se couper en un seul point, partager une droite (comme les pages d'un livre ouvert), ou n'avoir aucun point commun à tous les trois (comme les faces d'un prisme), même si deux d'entre eux se coupent toujours. D'autres configurations existent, avec des plans parallèles ou confondus.

Outil D.1 **Voir les plans en trois dimensions**

Les calculatrices 3D de GeoGebra et de Desmos tracent un plan dès qu'on entre son équation. Pour le système de l'exercice D.5, on entre les trois équations en remplaçant k et h par des curseurs. En déplaçant les curseurs, on voit le troisième plan pivoter : les trois plans se coupent en un point tant que $k \neq 5$, puis forment un prisme ou un livre ouvert lorsque $k = 5$, selon la valeur de h. Pour vérifier une réduction faite à la main, WolframAlpha calcule la forme échelonnée réduite d'une matrice augmentée : `row reduce {{1,2,1,1},{2,5,3,2},{-1,1,4,1}}` donne le résultat de l'exercice D.4. Il ne fournit toutefois pas les opérations effectuées : il permet de contrôler une réponse, pas de remplacer la démarche.

D.5.2 La méthode de Gauss–Jordan

Pour résoudre un système, on travaille sur sa **matrice augmentée** $(A \mid \mathbf{b})$, à laquelle on applique des **opérations élémentaires sur les lignes** :

1. échanger deux lignes, $L_i \leftrightarrow L_j$;
2. multiplier une ligne par un scalaire non nul, $L_i \leftarrow cL_i$;

3. ajouter à une ligne un multiple d'une autre, $L_i \leftarrow L_i + cL_j$.

Ces opérations ne changent pas l'ensemble des solutions, puisqu'on peut toujours les défaire. Le but est d'arriver à une matrice **échelonnée réduite** : dans chaque ligne non nulle, le premier coefficient non nul (le **pivot**) vaut 1, il est le seul coefficient non nul de sa colonne, et chaque pivot est situé à droite de celui de la ligne précédente. Les lignes nulles, s'il y en a, sont placées en bas.

Algorithme D.1 **Méthode de Gauss–Jordan**

Partant de la matrice augmentée, on traite les colonnes de gauche à droite.

1. Dans la colonne courante, choisir une ligne (parmi celles qui ne contiennent pas encore de pivot) dont le coefficient est non nul, et l'amener en position par un échange. Si aucun coefficient ne convient, passer à la colonne suivante.
2. Diviser cette ligne par son premier coefficient, pour que le pivot vaille 1.
3. Soustraire des multiples de cette ligne à *toutes* les autres lignes, pour annuler le reste de la colonne.
4. Recommencer avec la colonne suivante.

La méthode de *Gauss* proprement dite s'arrête à une forme échelonnée (des zéros sous les pivots seulement) et termine par une remontée ; Gauss–Jordan fait aussi le ménage au-dessus des pivots, de sorte que la solution se lit directement.

Exemple D.4 **Une solution unique**

Résolvons

$$\begin{cases} x + y + z = 6 \\ 2x + y - z = 1 \\ x - y + 2z = 5. \end{cases}$$

On écrit la matrice augmentée et on enchaîne les opérations, indiquées à droite :

$$\left(\begin{array}{ccc|c} 1 & 1 & 1 & 6 \\ 2 & 1 & -1 & 1 \\ 1 & -1 & 2 & 5 \end{array}\right) \qquad \begin{array}{l} L_2 \leftarrow L_2 - 2L_1 \\ L_3 \leftarrow L_3 - L_1 \end{array}$$

$$\sim \left(\begin{array}{ccc|c} 1 & 1 & 1 & 6 \\ 0 & -1 & -3 & -11 \\ 0 & -2 & 1 & -1 \end{array}\right) \qquad \begin{array}{l} L_2 \leftarrow -L_2, \text{ puis} \\ L_1 \leftarrow L_1 - L_2,\ L_3 \leftarrow L_3 + 2L_2 \end{array}$$

$$\sim \left(\begin{array}{ccc|c} 1 & 0 & -2 & -5 \\ 0 & 1 & 3 & 11 \\ 0 & 0 & 7 & 21 \end{array}\right) \qquad \begin{array}{l} L_3 \leftarrow \tfrac{1}{7} L_3, \text{ puis} \\ L_1 \leftarrow L_1 + 2L_3,\ L_2 \leftarrow L_2 - 3L_3 \end{array}$$

$$\sim \left(\begin{array}{ccc|c} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 3 \end{array}\right).$$

La solution est $(x, y, z) = (1, 2, 3)$; on la vérifie en la substituant dans les trois équations. Les trois plans se coupent en un seul point.

Exemple D.5 **Une infinité de solutions, ou aucune**

Considérons maintenant

$$\begin{cases} x + y + 2z = 3 \\ x + 2y + 3z = 5 \\ 2x + 3y + 5z = 8. \end{cases}$$

Les opérations $L_2 \leftarrow L_2 - L_1$, $L_3 \leftarrow L_3 - 2L_1$, puis $L_3 \leftarrow L_3 - L_2$ et $L_1 \leftarrow L_1 - L_2$ mènent à

$$\left(\begin{array}{ccc|c} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{array}\right).$$

La dernière ligne, $0 = 0$, ne dit plus rien : la troisième équation était la somme des deux premières. La colonne de z ne contient pas de pivot ; z est une **variable libre**, que l'on remplace par un paramètre t. Les deux autres lignes donnent $x = 1 - t$ et $y = 2 - t$, soit

$$(x, y, z) = (1, 2, 0) + t\,(-1, -1, 1), \qquad t \in \mathbb{R}\,.$$

C'est l'équation paramétrique d'une droite de $\mathbb{R}^3$: les trois plans forment un « livre ouvert ». Si l'on remplace le second membre 8 par 9, les mêmes opérations mènent à la ligne $(0\ 0\ 0 \mid 1)$, c'est-à-dire à l'équation impossible $0 = 1$: le système n'a alors aucune solution.

Ces deux exemples illustrent la règle générale. Une fois la matrice augmentée échelonnée réduite, le système est *incompatible* si une ligne de la forme $(0 \cdots 0 \mid c)$, avec $c \neq 0$, apparaît ; sinon, il a une solution unique s'il n'y a pas de variable libre, et une infinité de solutions, décrites par autant de paramètres qu'il y a de variables libres, dans le cas contraire. Le nombre de pivots de la forme échelonnée de A s'appelle le **rang** de A, noté $\operatorname{rang} A$.

Pour la classe D.2 **La méthode de réduction, en plus grand**

Au secondaire, les systèmes 2×2 se résolvent par comparaison, par substitution ou par *réduction*. Cette dernière, qui consiste à additionner des multiples des équations pour faire disparaître une inconnue, est exactement une opération élémentaire du type $L_i \leftarrow L_i + cL_j$. La méthode de Gauss–Jordan n'est donc rien d'autre que la réduction, appliquée de façon systématique et écrite sans les inconnues. Les trois cas de la figure D.4 (droites sécantes, parallèles ou confondues) sont eux aussi étudiés en troisième et quatrième secondaire ; on peut les faire apparaître graphiquement dans GeoGebra avant de les retrouver par le calcul.

Note historique D.1 **Une méthode bien plus ancienne que Gauss**

La méthode d'élimination figure déjà dans les *Neuf chapitres sur l'art du calcul*, un ouvrage chinois compilé il y a environ deux mille ans. Son huitième chapitre, *fangcheng* (« disposition rectangulaire »), résout des systèmes de plusieurs équations en disposant les coefficients en colonnes de bâtonnets sur une planche à calculer, puis en les combinant, ce qui revient à travailler sur la matrice augmentée. En Europe, Gauss utilise l'élimination au début du XIXe siècle pour ses calculs d'orbites (notamment celle de l'astéroïde Pallas) par la méthode des moindres

carrés. Le nom de Jordan est celui de Wilhelm Jordan, un géodésien allemand qui popularise la variante « réduite » en 1888 dans un manuel d'arpentage, et non celui du mathématicien Camille Jordan. Le nom « élimination de Gauss » lui-même ne s'impose qu'au xxe siècle [Grc11].

Pour des systèmes de taille raisonnable, on délègue ce travail à l'ordinateur. La bibliothèque `sympy` fournit la forme échelonnée réduite exacte, ce qui permet de vérifier un calcul fait à la main.

Programme D.1 : Gauss–Jordan, noyau et image avec `sympy` (exemple D.5).

```
import sympy as sp

M = sp.Matrix([[1, 1, 2],
               [1, 2, 3],
               [2, 3, 5]])
b = sp.Matrix([3, 5, 8])

R, pivots = M.row_join(b).rref()   # Gauss-Jordan sur [M | b]
print(R, pivots)
print("noyau :", M.nullspace())
print("image :", M.columnspace())
```

Programme D.2 : Sortie (mise en forme légèrement condensée).

```
Matrix([[1, 0, 1, 1], [0, 1, 1, 2], [0, 0, 0, 0]]) (0, 1)
noyau : [Matrix([[-1], [-1], [1]])]
image : [Matrix([[1], [1], [2]]), Matrix([[1], [2], [3]])]
```

Le tuple `(0, 1)` donne les colonnes des pivots (Python numérote à partir de 0). Nous reviendrons sur le noyau et l'image à la section D.7.

D.6 Indépendance linéaire, bases et dimension

Définition D.6 **Indépendance linéaire**

Les vecteurs $\mathbf{v}_1, \dots, \mathbf{v}_k$ sont **linéairement indépendants** si la seule combinaison linéaire nulle est la combinaison triviale :

$$c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k = \mathbf{0} \quad \Longrightarrow \quad c_1 = \cdots = c_k = 0.$$

Sinon, ils sont **linéairement dépendants**.

Dire que des vecteurs sont dépendants revient à dire que l'un d'eux est une combinaison linéaire des autres : il est « de trop », puisqu'il n'ajoute aucune direction nouvelle à l'espace engendré. Deux vecteurs sont dépendants exactement lorsqu'ils sont parallèles ; trois vecteurs de $\mathbb{R}^3$ le sont exactement lorsqu'ils sont dans un même plan passant par l'origine.

Pour tester l'indépendance, on place les vecteurs en colonnes dans une matrice A et on résout le système homogène $A\mathbf{c} = \mathbf{0}$: les vecteurs sont indépendants si et seulement si ce

système n'admet que la solution nulle, c'est-à-dire si chaque colonne contient un pivot. Par exemple, les colonnes $(1, 1, 2)$, $(1, 2, 3)$ et $(2, 3, 5)$ de la matrice de l'exemple D.5 sont dépendantes, car la troisième colonne n'a pas de pivot. La forme échelonnée donne même la relation : la troisième colonne est la somme des deux premières. Retenons aussi qu'on ne peut pas trouver plus de n vecteurs indépendants dans $\mathbb{R}^n$, puisqu'une matrice à n lignes a au plus n pivots.

Définition D.7 **Base, dimension, coordonnées**

Une **base** d'un sous-espace V est une famille de vecteurs linéairement indépendants qui engendrent V. Toutes les bases de V ont le même nombre d'éléments, appelé la **dimension** de V. Si $(\mathbf{b}_1, \dots, \mathbf{b}_k)$ est une base de V, tout $\mathbf{x} \in V$ s'écrit *de façon unique* $\mathbf{x} = c_1\mathbf{b}_1 + \cdots + c_k\mathbf{b}_k$; les nombres $c_1, \dots, c_k$ sont les **coordonnées** de $\mathbf{x}$ dans cette base.

La **base standard** $(\mathbf{e}_1, \dots, \mathbf{e}_n)$ de $\mathbb{R}^n$ est formée des colonnes de I_n ; dans cette base, les coordonnées d'un vecteur sont simplement ses composantes. Mais il existe une infinité d'autres bases, et changer de base revient à changer de quadrillage. Dans l'exemple D.1, nous avons trouvé que $(4, 7) = 1 \cdot \mathbf{u} + 2 \cdot \mathbf{v}$: dans la base $(\mathbf{u}, \mathbf{v})$, le vecteur $(4, 7)$ a pour coordonnées $(1, 2)$ (figure D.6).

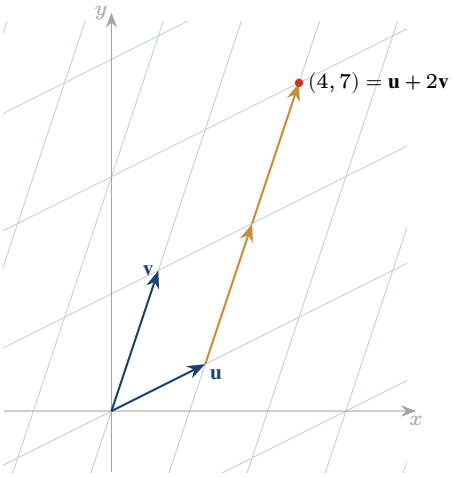


FIGURE D.6 : Le quadrillage associé à la base $(\mathbf{u}, \mathbf{v})$, avec $\mathbf{u} = (2, 1)$ et $\mathbf{v} = (1, 3)$. Pour atteindre le point $(4, 7)$, on avance d'un pas selon $\mathbf{u}$ et de deux pas selon $\mathbf{v}$: ses coordonnées dans cette base sont $(1, 2)$.

Dans $\mathbb{R}^n$, une famille de n vecteurs est une base dès qu'elle est indépendante, ou dès qu'elle engendre $\mathbb{R}^n$: une seule des deux conditions suffit. Les bases les plus commodes sont les **bases orthonormées**, formées de vecteurs unitaires deux à deux orthogonaux. Les coordonnées s'y calculent sans résoudre de système : la coordonnée de $\mathbf{x}$ selon $\mathbf{b}_i$ est simplement $\mathbf{x} \cdot \mathbf{b}_i$.

D.7 Noyau et image

Définition D.8 **Noyau, image**

Soit A une matrice $m \times n$. Le **noyau** de A est l'ensemble des vecteurs que A envoie sur $\mathbf{0}$:

$$\ker A = \{\mathbf{x} \in \mathbb{R}^n : A\mathbf{x} = \mathbf{0}\}.$$

L'**image** de A est l'ensemble des vecteurs atteints, $\operatorname{Im} A = \{A\mathbf{x} : \mathbf{x} \in \mathbb{R}^n\}$. D'après (D.2), c'est l'espace engendré par les colonnes de A.

Le noyau est un sous-espace de $\mathbb{R}^n$, l'image un sous-espace de $\mathbb{R}^m$. On trouve une base du noyau en résolvant $A\mathbf{x} = \mathbf{0}$ par Gauss–Jordan : chaque variable libre fournit un vecteur de base. Les colonnes de A qui contiennent un pivot forment, quant à elles, une base de l'image. Il y a donc autant de vecteurs dans une base de l'image que de pivots, et autant dans une base du noyau que de variables libres, d'où le **théorème du rang** :

$$\dim \ker A + \operatorname{rang} A = n. \tag{D.3}$$

Ce théorème a une lecture géométrique très parlante : ce que la matrice « écrase » (le noyau) et ce qu'elle « conserve » (l'image) se partagent les n dimensions de départ. Prenons $A = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$. Ses deux colonnes sont parallèles, donc son image est la droite engendrée par $(1, 2)$, de dimension 1. Son noyau est la droite engendrée par $(-2, 1)$, puisque $A \begin{pmatrix} -2 \\ 1 \end{pmatrix} = \mathbf{0}$. La matrice écrase le plan sur une droite : chaque droite parallèle au noyau est envoyée sur un seul point (figure D.7).

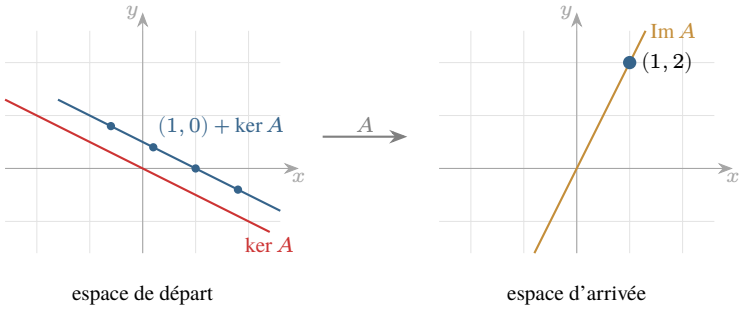


FIGURE D.7 : La matrice $A = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$ envoie tout le plan sur la droite Im A. Les points d'une même droite parallèle au noyau (en bleu) ont tous la même image, ici le point $(1, 2)$.

Le noyau mesure donc exactement le défaut d'injectivité : $A\mathbf{x} = A\mathbf{y}$ si et seulement si $\mathbf{x} - \mathbf{y} \in \ker A$. C'est l'argument utilisé à la proposition 1.2 du premier chapitre. Il en découle une description complète des solutions d'un système : *si* $\mathbf{x}_p$ *est une solution particulière de* $A\mathbf{x} = \mathbf{b}$, *toutes les solutions s'écrivent* $\mathbf{x}_p + \mathbf{h}$, *avec* $\mathbf{h} \in \ker A$. L'ensemble des solutions est un translaté du noyau. Dans l'exemple D.5, le noyau est la droite engendrée par $(-1, -1, 1)$, et les solutions forment la droite parallèle passant par $(1, 2, 0)$ (figure D.8).

D.8 Le déterminant

Le déterminant d'une matrice carrée est un nombre qui résume, en un seul coup, l'effet de la matrice sur les aires (en dimension 2) ou sur les volumes (en dimension 3). Pour

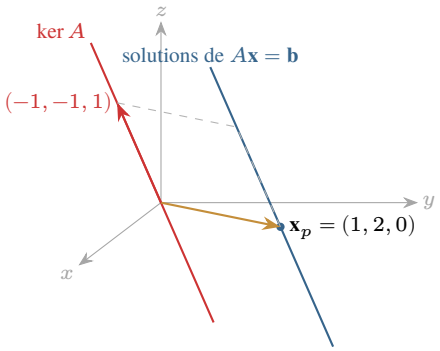


FIGURE D.8 : L'ensemble des solutions de l'exemple D.5 (en bleu) est la droite parallèle au noyau (en rouge) qui passe par une solution particulière $\mathbf{x}_p$.

une matrice 2×2,

$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc,$$

et nous avons vu à la proposition 1.4 que $|\det A|$ est l'aire du parallélogramme construit sur les colonnes de A, tandis que son signe indique si l'orientation est conservée.

Pour une matrice 3×3, on développe selon la première ligne :

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

Chaque coefficient de la ligne est multiplié par le déterminant de la matrice obtenue en supprimant sa ligne et sa colonne, avec des signes alternés. Cette **expansion par cofacteurs** se généralise à toute taille et peut se faire selon n'importe quelle ligne ou colonne (les signes suivent alors un damier $+, -, +, \ldots$ qui commence par $+$ en haut à gauche). Pour la matrice de l'exemple D.4,

$$\det \begin{pmatrix} 1 & 1 & 1 \\ 2 & 1 & -1 \\ 1 & -1 & 2 \end{pmatrix} = 1\,(2-1) - 1\,(4+1) + 1\,(-2-1) = -7.$$

L'interprétation géométrique se prolonge : $|\det A|$ est le volume du **parallélépipède** construit sur les trois colonnes de A (figure D.9), et le signe indique si ces colonnes, prises dans l'ordre, forment une base directe, c'est-à-dire qui respecte la « règle de la main droite ». Une application linéaire de matrice A multiplie donc tous les volumes par $|\det A|$.

Cette image rend presque évidentes les propriétés suivantes, que nous admettrons.

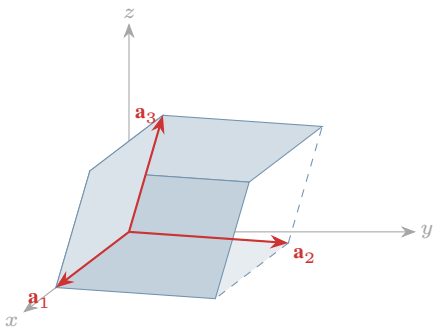


FIGURE D.9 : Le parallélépipède construit sur les colonnes $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$ d'une matrice 3×3 a pour volume $|\det A|$. Si les trois vecteurs sont dans un même plan, le solide est aplati et son volume est nul.

Proposition D.1 **Propriétés du déterminant**

Soient A et B deux matrices $n \times n$.

1. $\det A \neq 0$ si et seulement si les colonnes de A sont linéairement indépendantes.
2. $\det(AB) = \det(A)\ \det(B)$ et $\det(A^\intercal) = \det(A)$.
3. Le déterminant d'une matrice triangulaire est le produit des coefficients de sa diagonale.
4. Échanger deux lignes change le signe du déterminant ; multiplier une ligne par c multiplie le déterminant par c ; ajouter à une ligne un multiple d'une autre ne le change pas.

La propriété (1) est la plus importante : un déterminant nul signale un parallélépipède aplati, donc des colonnes dépendantes et un noyau non trivial. La propriété (2) traduit le fait que les facteurs de volume se multiplient quand on compose deux transformations. La propriété (4), enfin, fournit une méthode de calcul bien plus efficace que l'expansion par cofacteurs pour les grandes matrices : on échelonne par Gauss, en tenant compte des échanges de lignes, puis on multiplie les pivots. C'est ce que font les logiciels : l'expansion par cofacteurs d'une matrice $n \times n$ demande de l'ordre de $n!$ opérations, ce qui la rend inutilisable dès que n dépasse une quinzaine, alors que l'échelonnement en demande de l'ordre de n^3.

D.9 L'inverse d'une matrice

Définition D.9 **Matrice inversible**

Une matrice carrée A de taille n est **inversible** s'il existe une matrice B telle que $AB = BA = I_n$. Cette matrice est alors unique ; on l'appelle l'**inverse** de A et on la note A^{-1}.

Géométriquement, A^{-1} est la transformation qui défait ce que A a fait : si A tourne de 30°, A^{-1} tourne de $-30°$; si A étire d'un facteur 3 dans une direction, A^{-1} comprime d'un facteur 3 dans cette même direction. Une matrice qui écrase l'espace (un noyau non trivial, un déterminant nul) ne peut pas avoir d'inverse, puisqu'on ne peut pas savoir de quel point

venait l'image. Pour une matrice 2×2 de déterminant non nul, on vérifie directement que

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} : \tag{D.4}$$

on échange les coefficients de la diagonale, on change le signe des deux autres, et on divise par le déterminant.

Exemple D.6 **Résoudre un système avec l'inverse**

Soit $A = \begin{pmatrix} 2 & 1 \\ 5 & 3 \end{pmatrix}$, de déterminant $6 - 5 = 1$. Par la formule (D.4), $A^{-1} = \begin{pmatrix} 3 & -1 \\ -5 & 2 \end{pmatrix}$, et on vérifie que $AA^{-1} = I_2$. Le système $2x + y = 4$, $5x + 3y = 11$ s'écrit $A\mathbf{x} = \mathbf{b}$; en multipliant à gauche par A^{-1},

$$\mathbf{x} = A^{-1}\mathbf{b} = \begin{pmatrix} 3 & -1 \\ -5 & 2 \end{pmatrix} \begin{pmatrix} 4 \\ 11 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

Deux règles méritent d'être retenues : $(A^{-1})^{-1} = A$ et

$$(AB)^{-1} = B^{-1}A^{-1}.$$

L'ordre s'inverse, comme lorsqu'on s'habille : on enfile les chaussettes, puis les souliers, mais pour se déshabiller on retire d'abord les souliers. On retrouve ici la remarque 1.1 du premier chapitre.

Pour une matrice plus grande, on peut calculer l'inverse par Gauss–Jordan, en échelonnant la matrice $(A \mid I_n)$: si l'on parvient à $(I_n \mid B)$, alors $B = A^{-1}$; si un pivot manque, A n'est pas inversible. À la main, ce calcul devient vite fastidieux au-delà de la taille 3, et il n'apporte pas grand-chose de plus que la méthode elle-même. *En pratique, nous ferons ces calculs numériquement, avec Python.*

Programme D.3 : Déterminant, inverse et résolution avec `numpy` (exemple D.4).

```
import numpy as np

A = np.array([[1,  1,  1],
              [2,  1, -1],
              [1, -1,  2]], dtype=float)
b = np.array([6, 1, 5], dtype=float)

print("det(A) =", np.linalg.det(A))
print(np.linalg.inv(A))
x = np.linalg.solve(A, b)      # résout Ax = b sans calculer l'inverse
print("x =", x)
print("vérification :", A @ x - b)
```

Programme D.4 : Sortie.

```
det(A) = -6.999999999999999
[[-0.14285714  0.42857143  0.28571429]
 [ 0.71428571 -0.14285714 -0.42857143]
 [ 0.42857143 -0.28571429  0.14285714]]
x = [1. 2. 3.]
vérification : [0. 0. 0.]
```

On retrouve le déterminant -7, à une erreur d'arrondi près, et la solution $(1, 2, 3)$. Les coefficients de l'inverse sont des septièmes ($0{,}142857\ldots = \frac{1}{7}$), ce qui n'a rien d'étonnant pour une matrice de déterminant -7. Si l'on tente d'inverser la matrice de l'exemple D.5, `numpy` renvoie l'erreur `LinAlgError: Singular matrix`, et `np.linalg.matrix_rank` indique un rang de 2.

Remarque D.2 **Résoudre plutôt qu'inverser**

Pour résoudre $A\mathbf{x} = \mathbf{b}$, il est préférable d'appeler `np.linalg.solve(A, b)` plutôt que de calculer `np.linalg.inv(A) @ b`. La fonction `solve` applique directement une élimination de Gauss (sous la forme d'une factorisation LU), ce qui demande environ trois fois moins d'opérations et accumule moins d'erreurs d'arrondi. L'inverse reste utile comme objet théorique, ou lorsqu'on doit réellement l'examiner, mais on le calcule rarement pour lui-même.

Nous pouvons maintenant rassembler, en un seul énoncé, les liens entre les notions vues jusqu'ici. C'est probablement le résultat le plus utile de toute l'annexe.

Théorème D.2 **Caractérisations des matrices inversibles**

Pour une matrice carrée A de taille n, les énoncés suivants sont équivalents :

1. A est inversible ;
2. $\det A \neq 0$;
3. pour tout $\mathbf{b} \in \mathbb{R}^n$, le système $A\mathbf{x} = \mathbf{b}$ a une solution unique ;
4. $\ker A = \{\mathbf{0}\}$;
5. $\operatorname{rang} A = n$, c'est-à-dire que la forme échelonnée réduite de A est I_n ;
6. les colonnes de A forment une base de $\mathbb{R}^n$.

Géométriquement, tous ces énoncés disent la même chose : la transformation $\mathbf{x} \mapsto A\mathbf{x}$ n'écrase aucune direction.

D.10 Valeurs propres et vecteurs propres

Une matrice 2×2 quelconque déforme le plan de façon parfois difficile à visualiser : les vecteurs changent à la fois de longueur et de direction. Mais certaines directions privilégiées sont simplement étirées ou comprimées, sans tourner. Les trouver, c'est souvent comprendre la transformation tout entière.

Définition D.10 **Valeur propre, vecteur propre**

Soit A une matrice carrée. Un nombre λ est une **valeur propre** de A s'il existe un vecteur *non nul* $\mathbf{v}$ tel que

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

Un tel vecteur $\mathbf{v}$ est un **vecteur propre** associé à λ. Les vecteurs propres associés à λ, auxquels on ajoute $\mathbf{0}$, forment un sous-espace appelé **sous-espace propre**.

Un vecteur propre est donc envoyé sur un multiple de lui-même : la droite qu'il engendre est *invariante* par la transformation. La valeur propre indique ce qui se passe le long de

cette droite : un étirement si $|\lambda| > 1$, une compression si $|\lambda| < 1$, un retournement si $\lambda < 0$, et un écrasement sur l'origine si $\lambda = 0$. En particulier, 0 est une valeur propre de A exactement lorsque $\ker A \neq \{\mathbf{0}\}$, ce qui ajoute un septième énoncé au théorème D.2 : *A est inversible si et seulement si 0 n'est pas une valeur propre de A.*

D.10.1 Le polynôme caractéristique

L'équation $A\mathbf{v} = \lambda\mathbf{v}$ s'écrit $(A - \lambda I)\mathbf{v} = \mathbf{0}$. Elle a une solution non nulle exactement lorsque la matrice $A - \lambda I$ n'est pas inversible, c'est-à-dire lorsque

$$\det(A - \lambda I) = 0.$$

Le membre de gauche est un polynôme de degré n en λ, appelé **polynôme caractéristique** de A ; ses racines sont les valeurs propres. Une fois une valeur propre trouvée, on obtient les vecteurs propres associés en calculant le noyau de $A - \lambda I$, par Gauss–Jordan.

Pour une matrice 2×2, le calcul donne $\det(A - \lambda I) = \lambda^2 - (a + d)\lambda + (ad - bc)$. La somme $a + d$ des coefficients diagonaux s'appelle la **trace** de A ; elle est égale à la somme des valeurs propres, tandis que le déterminant en est le produit. C'est une vérification rapide, qui évite bien des erreurs de calcul.

Exemple D.7 **Un étirement caché**

Soit $A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$. Son polynôme caractéristique est

$$\det \begin{pmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{pmatrix} = (2-\lambda)^2 - 1 = (\lambda - 1)(\lambda - 3),$$

et les valeurs propres sont $\lambda_1 = 3$ et $\lambda_2 = 1$ (leur somme vaut bien la trace 4, et leur produit le déterminant 3). Pour $\lambda_1 = 3$, le noyau de $A - 3I = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$ est engendré par $\mathbf{v}_1 = (1, 1)$; pour $\lambda_2 = 1$, celui de $A - I = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ est engendré par $\mathbf{v}_2 = (1, -1)$.
Géométriquement, A étire le plan d'un facteur 3 le long de la diagonale $y = x$ et laisse inchangée la direction perpendiculaire. Le cercle unité devient une ellipse dont les axes sont portés par les vecteurs propres (figure D.10). Un vecteur qui n'est pas propre, comme $\mathbf{e}_1$, change de direction : il est « attiré » vers la direction la plus étirée.

Les transformations du chapitre 1 fournissent des exemples qu'on peut deviner sans calcul. Une réflexion par rapport à une droite passant par l'origine a pour valeurs propres 1 (les vecteurs de l'axe, qui ne bougent pas) et -1 (les vecteurs perpendiculaires à l'axe, qui sont retournés). Une homothétie de rapport k a une seule valeur propre, k, et tous les vecteurs non nuls sont propres. Une rotation d'angle θ strictement compris entre 0 et π ne laisse aucune direction invariante : elle n'a aucune valeur propre réelle, et son polynôme caractéristique $\lambda^2 - 2\cos\theta\,\lambda + 1$ a un discriminant négatif. Ses valeurs propres sont les nombres complexes $e^{\pm i\theta}$, ce qui explique que Python affiche parfois des valeurs propres complexes. Enfin, le cisaillement $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ a pour seule valeur propre 1, mais ses vecteurs propres ne remplissent qu'une droite, l'axe des x : nous verrons que c'est un obstacle à la diagonalisation.

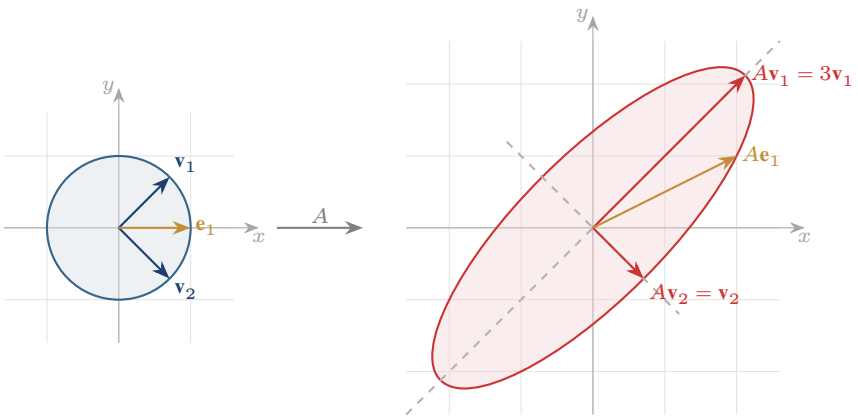


FIGURE D.10 : La matrice $A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$ transforme le cercle unité en une ellipse. Les vecteurs propres $\mathbf{v}_1$ et $\mathbf{v}_2$ (ici normalisés) gardent leur direction : ils sont étirés d'un facteur 3 et 1 respectivement. Le vecteur $\mathbf{e}_1$, qui n'est pas propre, tourne vers la direction de $\mathbf{v}_1$.

D.11 Diagonalisation

Les matrices les plus simples sont les matrices **diagonales** : la matrice $D = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$ étire l'axe des x d'un facteur λ_1 et l'axe des y d'un facteur λ_2. C'est une contraction ou une dilatation non uniforme au sens de la définition 1.10, et ses puissances se calculent sans effort : $D^k = \begin{pmatrix} \lambda_1^k & 0 \\ 0 & \lambda_2^k \end{pmatrix}$. Diagonaliser une matrice, c'est montrer qu'elle se comporte comme une matrice diagonale, pourvu qu'on la regarde dans une base bien choisie, formée de vecteurs propres.

Définition D.11 **Matrice diagonalisable**

Une matrice carrée A de taille n est **diagonalisable** s'il existe une matrice inversible P et une matrice diagonale D telles que

$$A = PDP^{-1}.$$

Théorème D.3 **Critère de diagonalisation**

Une matrice A de taille n est diagonalisable si et seulement si elle possède n vecteurs propres linéairement indépendants. Dans ce cas, on peut prendre pour P la matrice dont les colonnes sont ces vecteurs propres, et pour D la matrice diagonale des valeurs propres correspondantes, dans le même ordre. C'est en particulier le cas si A a n valeurs propres distinctes.

L'égalité $A = PDP^{-1}$ se lit de droite à gauche comme une recette en trois étapes (figure D.11) : P^{-1} calcule les coordonnées du vecteur dans la base propre ; D étire chaque coordonnée par la valeur propre correspondante ; P revient enfin à la base standard. Autrement dit, *une matrice diagonalisable n'est qu'un étirement non uniforme, vu dans une base oblique*. Le cisaillement $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, qui n'a qu'une droite de vecteurs propres, n'est pas diagonalisable : aucun choix de base ne le transforme en un simple étirement.

Exemple D.8 **Diagonaliser et calculer une puissance**

Pour $A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$, l'exemple D.7 fournit

$$P = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \qquad D = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}, \qquad P^{-1} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

On peut vérifier que $PDP^{-1} = A$. L'intérêt apparaît quand on calcule des puissances : les facteurs $P^{-1}P$ se simplifient au milieu, de sorte que

$$A^k = (PDP^{-1})(PDP^{-1})\cdots(PDP^{-1}) = PD^kP^{-1} = \frac{1}{2}\begin{pmatrix} 3^k+1 & 3^k-1 \\ 3^k-1 & 3^k+1 \end{pmatrix}.$$

Pour $k = 10$, on obtient $A^{10} = \begin{pmatrix} 29\,525 & 29\,524 \\ 29\,524 & 29\,525 \end{pmatrix}$ sans effectuer neuf produits matriciels.

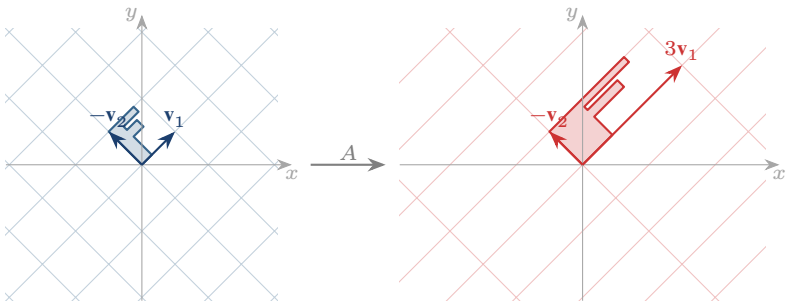


FIGURE D.11 : Dans la base propre (vecteurs normalisés), l'action de $A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$ se réduit à un étirement d'un facteur 3 dans la direction de $\mathbf{v}_1$, sans aucune déformation dans la direction perpendiculaire. Le quadrillage oblique reste un quadrillage, et le « F » est simplement étiré.

D.11.1 Le cas des matrices symétriques

La matrice de l'exemple D.8 est symétrique, et ses vecteurs propres $(1, 1)$ et $(1, -1)$ sont orthogonaux. Ce n'est pas un hasard.

Théorème D.4 **Théorème spectral**

Toute matrice symétrique réelle A est diagonalisable, ses valeurs propres sont réelles, et l'on peut choisir ses vecteurs propres de façon à former une base orthonormée. Autrement dit, il existe une matrice Q dont les colonnes forment une base orthonormée, et une matrice diagonale D, telles que

$$A = QDQ^{\mathsf{T}}.$$

Une matrice Q dont les colonnes forment une base orthonormée vérifie $Q^{\mathsf{T}}Q = I$, donc $Q^{-1} = Q^{\mathsf{T}}$: on l'appelle une matrice **orthogonale**. C'est la matrice d'une rotation ou d'une réflexion, c'est-à-dire d'une isométrie. Le théorème spectral affirme donc qu'une matrice symétrique est un étirement non uniforme le long de directions *perpendiculaires*. On comprend mieux ainsi la remarque 3.4 du chapitre 3 : pour une matrice symétrique, le plus grand facteur d'allongement est la plus grande valeur absolue d'une valeur propre. Les

matrices symétriques apparaissent partout en pratique : ce sont notamment les matrices de covariance en statistique et les matrices d'inertie en mécanique. Le théorème spectral est ainsi l'un des plus utilisés de l'algèbre linéaire.

D.11.2 Diagonaliser avec Python

Le calcul des valeurs propres par le polynôme caractéristique ne se fait bien à la main qu'en dimension 2, parfois 3. Au-delà, trouver les racines d'un polynôme de degré élevé est un problème numérique délicat, et les logiciels utilisent d'autres algorithmes, itératifs. En `numpy`, la fonction `np.linalg.eig` traite une matrice quelconque, et `np.linalg.eigh` une matrice symétrique.

Programme D.5 : Valeurs propres et diagonalisation avec `numpy` (exemple D.8).

```
import numpy as np

A = np.array([[2., 1.],
              [1., 2.]])
valeurs, P = np.linalg.eigh(A)  # eigh : pour une matrice symétrique
print("valeurs propres :", valeurs)
print(P)                        # les vecteurs propres, en colonnes

D = np.diag(valeurs)
print(np.allclose(P @ D @ P.T, A))       # A = P D P^T ?
print(P @ np.diag(valeurs**10) @ P.T)    # A^10, par diagonalisation
```

Programme D.6 : Sortie.

```
valeurs propres : [1. 3.]
[[-0.70710678  0.70710678]
 [ 0.70710678  0.70710678]]
True
[[29525. 29524.]
 [29524. 29525.]]
```

Trois observations s'imposent. D'abord, `numpy` range les valeurs propres dans l'ordre croissant et renvoie des vecteurs propres *unitaires* : la seconde colonne est $\frac{1}{\sqrt{2}}(1,1)$, et la première est $\frac{1}{\sqrt{2}}(-1,1)$, un multiple de notre $(1,-1)$. Un vecteur propre n'est défini qu'à un facteur non nul près ; il ne faut donc pas s'inquiéter si le logiciel n'affiche pas les mêmes vecteurs que le calcul à la main. Ensuite, la fonction `np.allclose` compare deux tableaux en tolérant les erreurs d'arrondi, ce qui est la bonne façon de tester une égalité en calcul numérique. Enfin, pour une matrice non symétrique, la fonction `np.linalg.eig` peut renvoyer des nombres complexes, affichés avec un suffixe `j` (par exemple `3.+0.j`), même lorsque leur partie imaginaire est nulle.

Exercices

Exercice D.1. Soit $\mathbf{u} = (1,-2,2)$ et $\mathbf{v} = (3,0,-4)$.

a) Calculer $2\mathbf{u} - \mathbf{v}$, $\|\mathbf{u}\|$ et $\|\mathbf{v}\|$.

b) Calculer $\mathbf{u} \cdot \mathbf{v}$ et l'angle entre $\mathbf{u}$ et $\mathbf{v}$. Cet angle est-il aigu ou obtus ?

c) Donner un vecteur unitaire de même direction et de même sens que $\mathbf{v}$.

d) Dans le plan, calculer la projection orthogonale du point $P = (7, 1)$ sur la droite passant par l'origine et dirigée par $(3, 4)$, puis la distance de P à cette droite. Faire un dessin.

Exercice D.2. Soit $\mathbf{u} = (1, 2, 1)$ et $\mathbf{v} = (2, 1, 0)$.

a) Écrire $\mathbf{w}_1 = (4, 5, 2)$ comme combinaison linéaire de $\mathbf{u}$ et de $\mathbf{v}$.

b) Montrer que $\mathbf{w}_2 = (1, 1, 1)$ n'est pas une combinaison linéaire de $\mathbf{u}$ et de $\mathbf{v}$.

c) Interpréter géométriquement les résultats précédents.

Exercice D.3. Soit $A = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & -1 \end{pmatrix}$ et $B = \begin{pmatrix} 1 & 0 \\ 2 & 1 \\ 0 & 3 \end{pmatrix}$.

a) Calculer AB et BA. Quelles sont leurs tailles ?

b) Vérifier que $(AB)^\intercal = B^\intercal A^\intercal$.

c) Soit $R = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. Calculer R^2 et R^4, et interpréter géométriquement.

Exercice D.4. Résoudre par la méthode de Gauss–Jordan, en indiquant chaque opération élémentaire :

$$\begin{cases} x + 2y + z = 1 \\ 2x + 5y + 3z = 2 \\ -x + y + 4z = 1. \end{cases}$$

Exercice D.5. On considère le système

$$\begin{cases} x + y + z = 1 \\ x + 2y + 3z = 2 \\ x + 3y + kz = h, \end{cases}$$

où k et h sont des paramètres réels.

a) Pour quelles valeurs de k et de h le système a-t-il une solution unique ? une infinité de solutions ? aucune solution ?

b) Dans le cas d'une infinité de solutions, les décrire et donner leur interprétation géométrique.

Exercice D.6. Les familles suivantes sont-elles linéairement indépendantes ? Si ce n'est pas le cas, donner une relation de dépendance.

a) $(1, 2, 3)$, $(4, 5, 6)$, $(7, 8, 9)$.

b) $(1, 0, 1)$, $(1, 1, 0)$, $(0, 1, 1)$.

c) $(1, 2, 0)$, $(0, 1, 5)$, $(2, -1, 3)$, $(4, 4, 4)$.

Exercice D.7.

a) Montrer que $\mathbf{b}_1 = (1, 1)$ et $\mathbf{b}_2 = (1, -1)$ forment une base de $\mathbb{R}^2$, puis trouver les coordonnées de $(3, 5)$ dans cette base de deux façons : en résolvant un système, puis à l'aide de produits scalaires (en tenant compte du fait que la base est orthogonale, mais pas orthonormée).

b) Trouver une base du plan d'équation $x - y + 2z = 0$ dans $\mathbb{R}^3$. Quelle est sa dimension ?

Exercice D.8. Soit $A = \begin{pmatrix} 1 & 2 & -1 \\ 2 & 4 & 1 \\ 3 & 6 & 0 \end{pmatrix}$.

a) Trouver une base du noyau et une base de l'image de A, et vérifier le théorème du rang.

b) Montrer que l'image de A est le plan d'équation $x + y - z = 0$. Les vecteurs $(1, 1, 2)$ et $(1, 0, 0)$ sont-ils dans l'image ?

c) Résoudre $A\mathbf{x} = (1, 1, 2)$ et décrire géométriquement l'ensemble des solutions.

Exercice D.9.

a) Calculer le déterminant de $M_k = \begin{pmatrix} k & 1 & 1 \\ 1 & k & 1 \\ 1 & 1 & k \end{pmatrix}$ et déterminer pour quelles valeurs de k cette matrice est inversible.

b) Pour $k = 1$, décrire l'image de M_1. Pourquoi est-il cohérent que son déterminant soit nul ?

c) Calculer le volume du parallélépipède construit sur $(2, 0, 0)$, $(1, 3, 0)$ et $(1, 1, 4)$.

d) Si A est une matrice 3×3 de déterminant 5, que valent $\det(2A)$, $\det(A^{-1})$ et $\det(A^{\mathsf{T}}A)$?

Exercice D.10.

a) Calculer l'inverse de $A = \begin{pmatrix} 3 & 2 \\ 4 & 3 \end{pmatrix}$ et l'utiliser pour résoudre $3x + 2y = 1$, $4x + 3y = 2$.

b) Pour quelles valeurs de a la matrice $\begin{pmatrix} a & 2 \\ 2 & a \end{pmatrix}$ est-elle inversible ?

c) Soit R_θ la matrice de la rotation d'angle θ (définition 1.7). Montrer que $R_\theta^{-1} = R_\theta^{\mathsf{T}} = R_{-\theta}$ et interpréter.

Exercice D.11.

a) Sans calcul, trouver les valeurs propres et les vecteurs propres de la réflexion d'axe $y = x$, de matrice $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, puis vérifier par le calcul.

b) Trouver les valeurs propres et les vecteurs propres de $A = \begin{pmatrix} 4 & 1 \\ 2 & 3 \end{pmatrix}$. Vérifier avec la trace et le déterminant.

c) Montrer que $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ n'a qu'une valeur propre, et que ses vecteurs propres forment une seule droite.

Exercice D.12.

a) Diagonaliser la matrice $A = \begin{pmatrix} 4 & 1 \\ 2 & 3 \end{pmatrix}$ de l'exercice D.11, puis donner une formule pour A^k.

b) Soit $S = \begin{pmatrix} 3 & 4 \\ 4 & -3 \end{pmatrix}$. Trouver une matrice orthogonale Q et une matrice diagonale D telles que $S = QDQ^{\mathsf{T}}$. Montrer que $\frac{1}{5}S$ est une réflexion, dont on précisera l'axe.

Exercice D.13. On considère le système $A\mathbf{x} = \mathbf{b}$ avec

$$A = \begin{pmatrix} 2 & 1 & -1 & 3 \\ 1 & 3 & 2 & -1 \\ 4 & -1 & 1 & 2 \\ 3 & 2 & -2 & 1 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 8 \\ -1 \\ 13 \\ 0 \end{pmatrix}.$$

a) Calculer $\det A$ avec `numpy`. Que peut-on en conclure sur le nombre de solutions ?

b) Résoudre le système avec `np.linalg.solve`, puis vérifier le résultat en calculant le résidu $A\mathbf{x} - \mathbf{b}$.

c) Comparer avec `np.linalg.inv(A) @ b` et avec la solution exacte fournie par `sympy` (méthode `rref` appliquée à la matrice augmentée).

Exercice D.14. On considère le système de trois équations à quatre inconnues

$$\begin{cases} x_1 + 2x_2 - x_3 + x_4 = 2 \\ 2x_1 + 4x_2 + x_3 - x_4 = 1 \\ x_1 + 2x_2 + 2x_3 - 2x_4 = -1. \end{cases}$$

a) Avec `sympy`, calculer la forme échelonnée réduite de la matrice augmentée. Quelles sont les variables libres ?

b) Calculer une base du noyau de la matrice des coefficients avec `nullspace()`, ainsi que son rang, et vérifier le théorème du rang.

c) Écrire l'ensemble des solutions sous la forme $\mathbf{x}_p + \mathbf{h}$, avec $\mathbf{h} \in \ker A$. De quel type d'objet géométrique s'agit-il ?

Exercice D.15. Plusieurs notions de cette annexe ont un équivalent plus élémentaire au secondaire : les systèmes de deux équations à deux inconnues (troisième et quatrième secondaire) et les vecteurs (cinquième secondaire, séquence SN).

a) Au secondaire, les élèves résolvent un système d'équations par comparaison, par substitution ou par réduction. Laquelle de ces méthodes correspond à la méthode de Gauss–Jordan (section D.5) ? Quel avantage y a-t-il à ce que l'enseignant connaisse cette correspondance ?

b) Relier les trois cas possibles pour un système de deux équations à deux inconnues (solution unique, aucune solution, infinité de solutions) à la position relative de deux droites, puis au déterminant de la matrice des coefficients. Comment cette connaissance peut-elle guider le choix des exemples en classe ?

c) Le produit scalaire est au programme de la cinquième secondaire (SN). Proposer une situation, accessible aux élèves, où il intervient naturellement.

E Éléments de théorie des nombres

Cette annexe rappelle les principaux résultats d'arithmétique vus dans le cours d'éléments de mathématiques. Elle ne prétend pas en remplacer les notes. L'objectif est de disposer, en un seul endroit, des définitions, des énoncés et surtout des démonstrations, rédigées de façon suffisamment détaillée pour qu'on puisse les reprendre, les adapter ou s'en inspirer devant une classe du secondaire.

Toute l'annexe porte sur l'ensemble des entiers $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$. Nous partirons de la notion de divisibilité et de la division avec reste, que les élèves pratiquent depuis le primaire. Nous présenterons ensuite les nombres premiers, puis le plus grand commun diviseur et l'algorithme d'Euclide qui permet de le calculer. Le théorème de Bachet–Bézout et le lemme de Gauss qui en découle nous permettront de démontrer complètement le théorème fondamental de l'arithmétique : tout entier se décompose, et d'une seule façon, en produit de nombres premiers. La seconde moitié de l'annexe est consacrée aux congruences, l'« arithmétique de l'horloge » inventée par Gauss, et à trois de leurs résultats phares : le petit théorème de Fermat, le théorème d'Euler et le théorème des restes chinois.

L'ordre de présentation diffère un peu de l'ordre habituel des listes de matières (on y trouve souvent la factorisation première avant l'algorithme d'Euclide). Nous l'avons choisi pour que chaque démonstration ne s'appuie que sur des résultats déjà établis. La factorisation première est ainsi traitée en deux temps : son existence à la section E.2, et son unicité à la section E.5.

Ces notions, qui ont longtemps passé pour les plus « pures » des mathématiques, sont aujourd'hui au cœur de la technologie. Le chiffrement RSA, qui protège les échanges sur Internet, repose sur le théorème d'Euler (remarque E.9). Les chiffres de contrôle des codes ISBN, des cartes bancaires ou des numéros d'assurance sociale sont des calculs de congruences (exercice E.11). Les ordinateurs eux-mêmes calculent modulo 2^{32} ou 2^{64} lorsqu'ils additionnent des entiers.

E.1 Divisibilité et division euclidienne

Définition E.1 **Divisibilité**

Soient a et b deux entiers, avec $a \neq 0$. On dit que a **divise** b, et on écrit $a \mid b$, s'il existe un entier k tel que $b = ka$. On dit aussi que a est un **diviseur** de b, ou que b est un **multiple** de a. Lorsque a ne divise pas b, on écrit $a \nmid b$.

Par exemple, $3 \mid 12$ puisque $12 = 4 \times 3$, et $-5 \mid 35$ puisque $35 = (-7) \times (-5)$; en revanche, $4 \nmid 10$, car aucun entier k ne vérifie $10 = 4k$. Tout entier non nul divise 0, puisque $0 = 0 \times a$.

Remarque E.1 **Une relation, pas une opération**

Il ne faut pas confondre $a \mid b$, qui est une affirmation (vraie ou fausse), et a/b, qui est un nombre. Ainsi, « $3 \mid 12$ » est une phrase vraie, tandis que $12/3$ est le nombre 4. Cette confusion est fréquente chez les élèves comme chez les étudiants. Il est utile de lire systématiquement la barre verticale à voix haute : « 3 divise 12 ».

L'intérêt de la définition est qu'elle ramène une question sur la divisibilité à l'existence d'un entier k. Toutes les démonstrations de cette section suivent le même schéma : on écrit les hypothèses sous la forme « $b = ka$ », on calcule, puis on reconnaît dans le résultat un multiple de l'entier voulu.

Proposition E.1 **Propriétés de la divisibilité**

Soient a, b et c des entiers, avec $a \neq 0$.

(i) On a $a \mid a$, $a \mid 0$ et $1 \mid b$.

(ii) Si $a \mid b$, alors $a \mid bc$.

(iii) Si $a \mid b$ et $b \mid c$ (avec $b \neq 0$), alors $a \mid c$.

(iv) Si $a \mid b$ et $a \mid c$, alors $a \mid (bx + cy)$ pour tous les entiers x et y.

(v) Si $a \mid b$ et $b \neq 0$, alors $|a| \leq |b|$.

(vi) Si $a \mid b$ et $b \mid a$, alors $a = b$ ou $a = -b$.

Démonstration. Le point (i) découle de $a = 1 \times a$, $0 = 0 \times a$ et $b = b \times 1$. Les points (ii), (iii) et (vi) font l'objet de l'exercice E.1. Démontrons (iv) et (v).

Pour (iv), les hypothèses fournissent deux entiers k et j tels que $b = ka$ et $c = ja$. Pour tous les entiers x et y, on a donc

$$bx + cy = kax + jay = (kx + jy)\, a.$$

Puisque $kx + jy$ est un entier, $bx + cy$ est un multiple de a.

Pour (v), écrivons $b = ka$. Comme $b \neq 0$, l'entier k n'est pas nul, de sorte que $|k| \geq 1$. Il suit que $|b| = |k|\,|a| \geq |a|$. □

Le point (iv) sera utilisé constamment : un diviseur commun de deux nombres divise toutes leurs **combinaisons linéaires** (à coefficients entiers). En particulier, il divise leur somme et leur différence. Le point (v) dit qu'un entier non nul n'a qu'un nombre fini de diviseurs, tous compris entre $-|b|$ et $|b|$.

E.1.1 La division euclidienne

Lorsque a ne divise pas b, la division « ne tombe pas juste » ; il reste quelque chose. C'est la division avec reste que les élèves apprennent au primaire : 17 divisé par 5 donne 3, reste 2, car $17 = 3 \times 5 + 2$. Pour énoncer et démontrer ce résultat avec soin, nous aurons besoin d'une propriété des entiers naturels qui paraît évidente, mais sur laquelle reposent presque toutes les démonstrations d'arithmétique.

Remarque E.2 **Le principe du bon ordre**

Toute partie non vide de $\mathbb{N} = \{0, 1, 2, \dots\}$ possède un plus petit élément. Cette propriété est fausse dans $\mathbb{Z}$ (l'ensemble des entiers négatifs n'a pas de plus petit élément) ou dans $]0, 1[$ (il n'y a pas de plus petit réel strictement positif). Elle est équivalente au principe de récurrence. Nous l'utiliserons sous la forme suivante : si une propriété est fausse pour au moins un entier naturel, il existe un plus petit entier pour lequel elle est fausse. On l'appelle un **plus petit contre-exemple**.

Théorème E.2 **Division euclidienne**

Soient $a \in \mathbb{Z}$ et b un entier strictement positif. Il existe un unique couple d'entiers (q, r) tel que

$$a = bq + r \qquad \text{et} \qquad 0 \leq r < b.$$

L'entier q s'appelle le **quotient** et r le **reste** de la division euclidienne de a par b.

Démonstration. L'idée est simple : parmi les nombres a, $a - b$, $a - 2b$, $a - 3b$, …, et $a + b$, $a + 2b$, …, on retient le plus petit qui soit encore positif ou nul.

Existence. Considérons l'ensemble

$$S = \{\, a - kb \;:\; k \in \mathbb{Z} \text{ et } a - kb \geq 0 \,\}.$$

C'est une partie de $\mathbb{N}$. Elle n'est pas vide : pour $k = -|a|$, on a $a - kb = a + |a|\, b \geq a + |a| \geq 0$, puisque $b \geq 1$. Par le principe du bon ordre, S possède un plus petit élément, que nous notons r ; il s'écrit $r = a - qb$ pour un certain entier q. Par construction, $a = bq + r$ et $r \geq 0$. Supposons par l'absurde que $r \geq b$. Alors $r - b = a - (q + 1)b$ est un élément de S (il est positif ou nul) strictement plus petit que r, ce qui contredit le choix de r. Donc $r < b$.

Unicité. Supposons que $a = bq + r = bq' + r'$, avec $0 \leq r < b$ et $0 \leq r' < b$. En soustrayant, on obtient $b(q - q') = r' - r$. Le membre de droite est strictement compris entre $-b$ et b ; c'est aussi un multiple de b. Le seul multiple de b dans $]-b, b[$ est 0. Ainsi $r' = r$, puis $b(q - q') = 0$, d'où $q = q'$ puisque $b \neq 0$. □

Exemple E.1 **Avec un dividende négatif**

Pour diviser -47 par 6, il ne suffit pas de diviser 47 par 6 et de changer les signes : $-47 = 6 \times (-7) - 5$ ne convient pas, car le reste doit être positif ou nul. On descend d'un cran :

$$-47 = 6 \times (-8) + 1, \qquad 0 \leq 1 < 6.$$

Le quotient est -8 et le reste est 1. Géométriquement, les multiples de 6 découpent la droite numérique en intervalles de longueur 6, et q est le numéro de l'intervalle $[6q, 6q + 6[$ qui contient a ; ici, -47 est dans $[-48, -42[$.

On remarque que $b \mid a$ si et seulement si le reste de la division de a par b est nul. C'est ce critère qu'appliquent les calculatrices et les langages de programmation : en Python, `a % b` donne le reste et `a // b` le quotient, avec la même convention que ci-dessus lorsque $b > 0$ (`-47 % 6` vaut bien `1`).

E.1.2 Diviseurs communs et multiples communs

Définition E.2 **PGCD, PPCM, nombres premiers entre eux**

Soient a et b deux entiers qui ne sont pas tous les deux nuls. Le **plus grand commun diviseur** de a et de b, noté $\text{pgcd}(a,b)$, est le plus grand entier qui divise à la fois a et b. Lorsque $\text{pgcd}(a,b) = 1$, on dit que a et b sont **premiers entre eux** (ou **copremiers**).
Si a et b sont tous deux non nuls, leur **plus petit commun multiple**, noté $\text{ppcm}(a,b)$, est le plus petit entier strictement positif qui est à la fois un multiple de a et un multiple de b.

Le PGCD est bien défini : 1 est un diviseur commun, et si par exemple $a \neq 0$, tout diviseur commun est au plus $|a|$ par le point (v) de la proposition E.1. Il est donc toujours au moins égal à 1. Le PPCM est bien défini aussi, puisque $|ab|$ est un multiple commun strictement positif. On a par exemple $\text{pgcd}(12,18) = 6$, $\text{ppcm}(12,18) = 36$, $\text{pgcd}(8,15) = 1$ et $\text{pgcd}(a,0) = |a|$ pour tout $a \neq 0$. Les signes ne jouent aucun rôle : $\text{pgcd}(a,b) = \text{pgcd}(|a|,|b|)$. C'est pourquoi, dans la suite, nous travaillerons le plus souvent avec des entiers positifs.

Pour la classe E.1 **Le PGCD et le PPCM en classe**

Au primaire et au premier cycle du secondaire, les élèves rencontrent les diviseurs et les multiples, les nombres premiers, les critères de divisibilité par 2, 3, 5, 9 et 10 ainsi que la notation exponentielle. Le PGCD intervient lorsqu'on réduit une fraction à sa plus simple expression ($\frac{12}{18} = \frac{2}{3}$, en divisant le numérateur et le dénominateur par 6) et le PPCM lorsqu'on cherche le plus petit dénominateur commun ($\frac{1}{12} + \frac{1}{18} = \frac{3}{36} + \frac{2}{36}$). On les calcule alors en dressant la liste des diviseurs ou des multiples, ou à l'aide de la factorisation première. Les méthodes de cette annexe, en particulier l'algorithme d'Euclide, sont bien plus efficaces dès que les nombres deviennent grands, et elles expliquent pourquoi les méthodes scolaires fonctionnent.

E.2 Nombres premiers

Définition E.3 **Nombre premier**

Un entier $p \geq 2$ est **premier** si ses seuls diviseurs positifs sont 1 et p. Un entier $n \geq 2$ qui n'est pas premier est dit **composé** : il s'écrit alors $n = ab$ avec $1 < a < n$ et $1 < b < n$.

Les premiers nombres premiers sont $2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47$. Le nombre 1 n'est pas premier ; nous verrons à la remarque E.6 que cette convention n'a rien d'arbitraire.

Lemme E.3 **Tout entier a un diviseur premier**

Soit $n \geq 2$ un entier.

(i) Le plus petit diviseur de n strictement supérieur à 1 est un nombre premier. En particulier, n possède au moins un diviseur premier.

(ii) Si n est composé, son plus petit diviseur premier p vérifie $p \leq \sqrt{n}$.

Démonstration. (i) L'ensemble des diviseurs de n strictement supérieurs à 1 n'est pas vide, puisqu'il contient n. Par le principe du bon ordre, il possède un plus petit élément p. Supposons que p ne soit pas premier : il aurait alors un diviseur d avec $1 < d < p$. Comme $d \mid p$ et $p \mid n$, on aurait $d \mid n$ (point (iii) de la proposition E.1). Ainsi d serait un diviseur de n strictement supérieur à 1 et plus petit que p, ce qui est impossible. Donc p est premier.

(ii) Si n est composé, écrivons $n = pm$, où p est son plus petit diviseur supérieur à 1. Puisque n n'est pas premier, $p < n$ et donc $m > 1$. Alors m est lui aussi un diviseur de n supérieur à 1, donc $m \geq p$ par minimalité de p. On en tire $n = pm \geq p \cdot p = p^2$, c'est-à-dire $p \leq \sqrt{n}$.

□

Le point (ii) fournit un test de primalité : pour savoir si n est premier, il suffit de vérifier qu'aucun nombre premier $p \leq \sqrt{n}$ ne le divise. Par exemple, $\sqrt{97} < 10$ et ni 2, ni 3, ni 5, ni 7 ne divisent 97 ; c'est donc un nombre premier. Le même raisonnement est à la base du **crible d'Ératosthène**. Pour obtenir tous les nombres premiers jusqu'à N, on écrit les entiers de 2 à N ; on entoure 2 et on raye ses multiples ; on entoure le premier nombre non rayé (3) et on raye ses multiples ; et ainsi de suite. On peut s'arrêter dès que le nombre entouré dépasse $\sqrt{N}$: tous les nombres non rayés qui restent sont alors premiers.

E.2.1 Une infinité de nombres premiers

Les nombres premiers se raréfient à mesure qu'on avance : il y en a 25 entre 1 et 100, mais seulement 16 entre 1001 et 1100. On pourrait croire qu'ils finissent par disparaître. Le résultat suivant, démontré par Euclide il y a plus de deux mille ans, affirme qu'il n'en est rien.

Théorème E.4 **Euclide**

Il existe une infinité de nombres premiers.

Démonstration. Raisonnons par l'absurde et supposons qu'il n'y ait qu'un nombre fini de nombres premiers, que nous énumérons tous : $p_1, p_2, \ldots, p_k$. Formons le nombre

$$N = p_1 p_2 \cdots p_k + 1.$$

Comme $N \geq 2 + 1 \geq 2$, le lemme E.3 garantit que N possède un diviseur premier p. Puisque la liste $p_1, \ldots, p_k$ contient tous les nombres premiers, p est l'un d'eux, disons $p = p_i$. Alors p divise N, et p divise aussi le produit $p_1 p_2 \cdots p_k$, dont il est un des facteurs. Par le point (iv) de la proposition E.1, p divise la différence

$$N - p_1 p_2 \cdots p_k = 1.$$

Or aucun nombre premier ne divise 1, puisque $p \geq 2$. Cette contradiction montre que notre hypothèse était fausse : la liste des nombres premiers ne peut pas être finie. □

Remarque E.3 **Le nombre N n'est pas forcément premier**

On lit parfois que la démonstration d'Euclide « fabrique » un nouveau nombre premier N. C'est inexact : la démonstration montre seulement que N a un facteur premier qui n'est pas dans la liste. Avec les six premiers nombres premiers, on obtient

$$2 \cdot 3 \cdot 5 \cdot 7 \cdot 11 \cdot 13 + 1 = 30\,031 = 59 \times 509,$$

qui n'est pas premier ; mais ses facteurs 59 et 509 sont bien absents de la liste de départ. Il est aussi instructif de remarquer que la démonstration n'est pas une recette pour trouver « le » nombre premier suivant : elle ne fait qu'établir qu'on ne peut jamais les avoir tous.

Note historique E.1 **Euclide d'Alexandrie**



Euclide, tel que l'imaginait Juste de Gand vers 1474 (palais ducal d'Urbino). L'inscription « Euclidi Megaren » le confond avec le philosophe Euclide de Mégare. Domaine public, Wikimedia Commons.

On ne sait presque rien de la vie d'Euclide, sinon qu'il enseignait à Alexandrie vers 300 av. J.-C., sous le règne de Ptolémée Ier. Aucun portrait d'époque ne nous est parvenu ; celui-ci est imaginé, près de dix-huit siècles plus tard. Son œuvre, en revanche, est sans doute le manuel de mathématiques le plus lu de l'histoire. Les Éléments, en treize livres, rassemblent et organisent le savoir géométrique et arithmétique des Grecs à partir d'un petit nombre de définitions, de postulats et de notions communes [Euc94].

Les livres VII, VIII et IX sont consacrés à l'arithmétique. On y trouve la définition des nombres premiers et des nombres premiers entre eux, l'algorithme qui porte aujourd'hui le nom d'Euclide (livre VII, propositions 1 et 2), une forme du lemme qui sera appelé plus tard lemme de Gauss (livre VII, proposition 30) et la démonstration de l'infinitude des nombres premiers (livre IX, proposition 20). Euclide ne dit d'ailleurs pas « il y a une infinité de nombres premiers », notion que les Grecs évitaient : il affirme que « les nombres premiers sont plus nombreux que toute multitude de nombres premiers proposée ». Plus que ses résultats, c'est sa méthode qui a marqué les mathématiques : chaque énoncé y est démontré à partir des précédents. Pendant plus de deux mille ans, les Éléments ont servi de modèle à ce qu'on appelle une démonstration [ORc].

E.2.2 Existence de la factorisation première

Proposition E.5 **Existence de la factorisation**

Tout entier $n \geq 2$ peut s'écrire comme un produit de nombres premiers (un produit d'un seul facteur étant permis).

Démonstration. Supposons le contraire. Par le principe du bon ordre, il existe alors un plus petit entier $n \geq 2$ qui n'est pas un produit de nombres premiers. Cet entier n n'est pas premier, sinon il serait un produit d'un seul facteur premier. Il est donc composé : $n = ab$, avec $1 < a < n$ et $1 < b < n$. Comme a et b sont plus petits que n et au moins égaux à 2, ils s'écrivent, par minimalité de n, comme des produits de nombres premiers. En mettant ces deux produits bout à bout, on obtient une écriture de $n = ab$ comme produit de nombres premiers, ce qui est une contradiction. □

En pratique, on obtient la factorisation en divisant successivement par les nombres premiers $2, 3, 5, 7, \ldots$: par exemple, $360 = 2 \times 180 = 2^2 \times 90 = 2^3 \times 45 = 2^3 \times 3^2 \times 5$. Rien, à ce stade, ne garantit qu'une autre façon de procéder (commencer par $360 = 10 \times 36$, par exemple) mène au même résultat. Cette unicité, que tout le monde tient pour acquise, demande en fait un argument supplémentaire : le lemme de Gauss, que nous démontrerons à la section E.5.

E.3 Le PGCD et l'algorithme d'Euclide

Pour calculer $\mathrm{pgcd}(1071, 462)$, on pourrait énumérer les diviseurs des deux nombres, ou les factoriser. Ces méthodes deviennent vite impraticables : factoriser un nombre de quelques centaines de chiffres est hors de portée des ordinateurs actuels. L'algorithme d'Euclide calcule le PGCD sans jamais factoriser, à l'aide de divisions euclidiennes successives. Il repose sur l'observation suivante.

Lemme E.6 **Le PGCD ne change pas quand on remplace a par son reste**

Soient a et b des entiers, avec $b > 0$, et soit $a = bq + r$ la division euclidienne de a par b. Alors a et b ont exactement les mêmes diviseurs communs que b et r. En particulier,

$$\mathrm{pgcd}(a, b) = \mathrm{pgcd}(b, r).$$

Démonstration. Soit d un diviseur commun de a et de b. Par le point (iv) de la proposition E.1, d divise la combinaison $a - qb = r$; c'est donc un diviseur commun de b et de r. Réciproquement, si d divise b et r, il divise la combinaison $qb + r = a$, et c'est un diviseur commun de a et de b. Les deux ensembles de diviseurs communs sont donc égaux. Ils ont en particulier le même plus grand élément. □

En répétant ce remplacement, on obtient des nombres de plus en plus petits, sans changer le PGCD.

Algorithme E.1 **Algorithme d'Euclide**

Données : deux entiers $a \geq b > 0$.

1. Poser $r_0 = a$, $r_1 = b$ et $k = 1$.
2. Tant que $r_k \neq 0$: effectuer la division euclidienne de r_{k-1} par r_k, $r_{k-1} = q_k r_k + r_{k+1}$ avec $0 \leq r_{k+1} < r_k$, puis remplacer k par $k+1$.
3. Renvoyer le dernier reste non nul, r_{k-1}.

Théorème E.7 **Validité de l'algorithme d'Euclide**

Pour tous les entiers $a \geq b > 0$, l'algorithme E.1 s'arrête après un nombre fini d'étapes, et le dernier reste non nul qu'il renvoie est $\operatorname{pgcd}(a, b)$.

Démonstration. Arrêt. Les restes successifs vérifient $b = r_1 > r_2 > r_3 > \cdots \geq 0$. Une suite strictement décroissante d'entiers positifs ou nuls ne peut pas compter plus de $b+1$ termes : il existe donc un indice n pour lequel $r_n \neq 0$ et $r_{n+1} = 0$, et l'algorithme s'arrête.

Résultat. Par le lemme E.6, appliqué à chacune des divisions effectuées,

$$\operatorname{pgcd}(a, b) = \operatorname{pgcd}(r_0, r_1) = \operatorname{pgcd}(r_1, r_2) = \cdots = \operatorname{pgcd}(r_n, r_{n+1}) = \operatorname{pgcd}(r_n, 0) = r_n.$$

La dernière égalité vient de ce que le plus grand diviseur de $r_n > 0$ est r_n lui-même, et que r_n divise 0. □

Exemple E.2 **Le PGCD de 1071 et 462**

Les divisions successives donnent

$$\begin{aligned} 1071 &= 2 \times 462 + 147, \\ 462 &= 3 \times 147 + 21, \\ 147 &= 7 \times 21 + 0. \end{aligned}$$

Le dernier reste non nul est 21 : $\operatorname{pgcd}(1071, 462) = 21$. Trois divisions ont suffi, alors que l'énumération des diviseurs de 1071 en demande beaucoup plus. On vérifie que $1071 = 21 \times 51$ et $462 = 21 \times 22$, et que 51 et 22 sont bien premiers entre eux.

Remarque E.4 **Un algorithme très rapide**

En 1844, Gabriel Lamé a démontré que le nombre de divisions effectuées par l'algorithme d'Euclide ne dépasse jamais cinq fois le nombre de chiffres du plus petit des deux nombres. Le pire cas est atteint par deux nombres de Fibonacci consécutifs, comme 34 et 21, pour lesquels tous les quotients valent 1. Ainsi, le PGCD de deux nombres de 300 chiffres s'obtient en au plus 1500 divisions, soit une fraction de seconde pour un ordinateur, alors qu'aucun algorithme connu ne permet de les factoriser en un temps raisonnable. Ce contraste entre la facilité du PGCD et la difficulté de la factorisation est précisément ce qui rend possible le chiffrement RSA (remarque E.9).

E.3.1 Une interprétation géométrique : paver un rectangle

L'algorithme d'Euclide a une interprétation visuelle très parlante, qui en est d'ailleurs la forme la plus ancienne. Considérons un rectangle de dimensions $a \times b$, avec $a \geq b$, et cherchons le plus grand carré qui permette de le paver exactement, c'est-à-dire de le recouvrir avec des copies de ce carré, sans chevauchement ni découpe.

On commence par découper dans le rectangle autant de carrés de côté b que possible. On en place q_1, où q_1 est le quotient de a par b, et il reste une bande de dimensions $r_2 \times b$, où r_2 est le reste. Dans cette bande, on découpe à nouveau le plus possible de carrés de côté r_2 ; il reste une bande plus petite. On recommence jusqu'à ce que la dernière bande soit entièrement remplie de carrés. Chaque étape correspond exactement à une division de l'algorithme d'Euclide, et le côté des derniers carrés est le dernier reste non nul, c'est-à-dire le PGCD.

La figure E.1 illustre le cas d'un rectangle 42×30. On y découpe un carré 30×30 ($42 = 1 \times 30 + 12$), puis deux carrés 12×12 dans la bande restante ($30 = 2 \times 12 + 6$), puis deux carrés 6×6 ($12 = 2 \times 6$), qui remplissent exactement ce qui reste. Le PGCD est 6, et le rectangle tout entier se pave avec $7 \times 5 = 35$ carrés de côté 6.

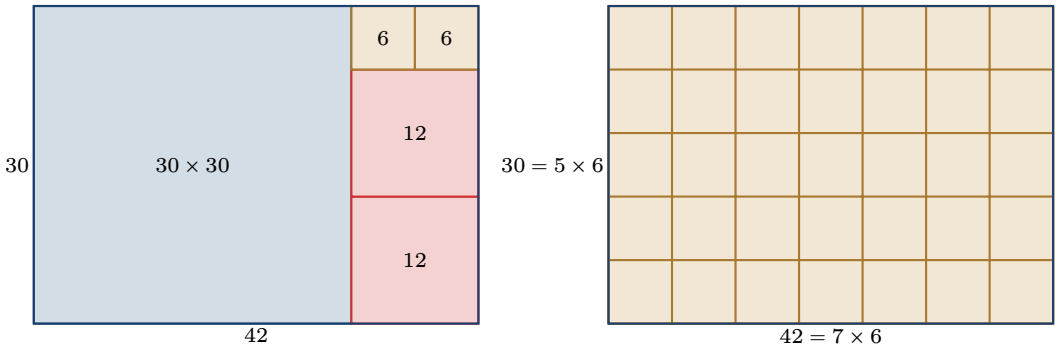


FIGURE E.1 : L'algorithme d'Euclide appliqué au rectangle 42×30. À gauche, on découpe successivement les plus grands carrés possibles : un carré de côté 30, deux de côté 12, puis deux de côté 6, qui remplissent la dernière bande. À droite, les carrés de côté $6 = \text{pgcd}(42, 30)$ pavent le rectangle entier ; ce sont les plus grands carrés qui le permettent.

Proposition E.8 **Le PGCD comme plus grand carré pavant**

Soient a et b des entiers strictement positifs. Un rectangle $a \times b$ peut être pavé par des carrés identiques de côté entier c si et seulement si c est un diviseur commun de a et de b. Le plus grand carré qui permet de le paver a donc pour côté $\text{pgcd}(a, b)$.

Démonstration. Si $c \mid a$ et $c \mid b$, on place a/c carrés le long du côté de longueur a et b/c carrés le long du côté de longueur b ; les $(a/c)(b/c)$ carrés ainsi disposés en grille pavent le rectangle.

Réciproquement, supposons que des carrés de côté c pavent le rectangle. Considérons le côté inférieur, de longueur a. Les carrés qui touchent ce côté sont posés côte à côte et le recouvrent exactement, sans chevaucher : s'il y en a m, alors $a = mc$, donc $c \mid a$. Le même argument appliqué au côté gauche montre que $c \mid b$. Les côtés possibles sont donc exactement les diviseurs communs, et le plus grand est le PGCD. □

Cette interprétation fournit aussi une justification visuelle du lemme E.6 : un carré qui pave le rectangle de départ pave chacun des carrés découpés (son côté divise b) et donc aussi la bande restante ; inversement, un carré qui pave la bande restante et les carrés découpés pave le rectangle entier.

Note historique E.2 Le plus vieil algorithme encore en usage

Euclide ne parle pas de divisions, mais de soustractions répétées : « le plus petit nombre étant continuellement retranché du plus grand » (livre VII, proposition 2). Les Grecs appelaient ce procédé anthyphérèse, littéralement « soustraction réciproque », et l'appliquaient aussi bien aux nombres qu'aux longueurs. C'est exactement le découpage de carrés de la figure E.1. Appliqué à deux longueurs comme le côté et la diagonale d'un carré, le procédé ne s'arrête jamais. C'est l'une des façons dont les Grecs ont compris que ces deux longueurs n'ont pas de commune mesure, autrement dit que $\sqrt{2}$ est irrationnel.
Donald Knuth, l'un des fondateurs de l'informatique théorique, voit dans l'algorithme d'Euclide le doyen de tous les algorithmes : c'est le plus ancien algorithme non trivial qui soit encore utilisé aujourd'hui [Knu97]. Il tourne en effet, sous une forme à peine modifiée, des milliards de fois par jour dans les navigateurs Web et les cartes à puce, chaque fois qu'une connexion sécurisée calcule un inverse modulaire (section E.6).

E.4 Le théorème de Bachet–Bézout

En « remontant » les calculs de l'algorithme d'Euclide, on peut écrire le PGCD de a et de b comme une combinaison de a et de b. Ce résultat, d'apparence technique, est la clé de presque tout ce qui suit.

Théorème E.9 Bachet–Bézout

Soient a et b deux entiers non tous les deux nuls, et $d = \operatorname{pgcd}(a, b)$. Il existe des entiers u et v tels que

$$au + bv = d.$$

Démonstration. Comme $\operatorname{pgcd}(a, b) = \operatorname{pgcd}(|a|, |b|)$ et qu'il suffit de changer le signe de u ou de v pour passer de a à $-a$, ou de b à $-b$, on peut supposer $a \geq b \geq 0$. Si $b = 0$, alors $d = a$ et l'on prend $u = 1$, $v = 0$. Supposons donc $b > 0$, et reprenons les notations de l'algorithme E.1 : $r_0 = a$, $r_1 = b$ et $r_{k+1} = r_{k-1} - q_k r_k$, jusqu'au dernier reste non nul $r_n = d$.

Montrons que chaque reste r_k est une combinaison de a et de b, c'est-à-dire qu'il s'écrit $r_k = au_k + bv_k$ pour des entiers u_k et v_k. C'est vrai pour $r_0 = a = a \cdot 1 + b \cdot 0$ et pour $r_1 = b = a \cdot 0 + b \cdot 1$. Supposons-le vrai pour deux restes consécutifs r_{k-1} et r_k. Alors

$$r_{k+1} = r_{k-1} - q_k r_k = (au_{k-1} + bv_{k-1}) - q_k(au_k + bv_k) = a\,(u_{k-1} - q_k u_k) + b\,(v_{k-1} - q_k v_k),$$

et r_{k+1} est aussi une combinaison de a et de b. De proche en proche, la propriété est vraie pour tous les restes, en particulier pour $r_n = d$. □

La démonstration est constructive : elle fournit un moyen de calculer u et v, appelé **algorithme d'Euclide étendu**. On dispose les calculs en lignes de la forme $r = au +$

bv, qu'on combine comme les lignes d'une matrice dans la méthode de Gauss–Jordan (section D.5). Chaque nouvelle ligne est l'avant-dernière moins le quotient fois la dernière.

Exemple E.3 **Coefficients de Bézout pour 1071 et 462**

Reprenons l'exemple E.2. Le tableau E.1 donne $21 = 1071 \times (-3) + 462 \times 7$. On le vérifie : $-3213 + 3234 = 21$.
On peut aussi remonter les divisions à la main, en partant de l'avant-dernière :

$$\begin{aligned} 21 &= 462 - 3 \times 147 \\ &= 462 - 3 \times (1071 - 2 \times 462) \\ &= 7 \times 462 - 3 \times 1071. \end{aligned}$$

Les coefficients ne sont pas uniques : la dernière ligne du tableau montre que $1071 \times 22 - 462 \times 51 = 0$. On peut donc ajouter 22 à u et retrancher 51 de v autant de fois qu'on le veut : $21 = 1071 \times 19 + 462 \times (-44)$, par exemple.

Table E.1 : L'algorithme d'Euclide étendu pour $a = 1071$ et $b = 462$. Chaque ligne vérifie $r = 1071\,u + 462\,v$; l'opération indiquée dans la première colonne utilise le quotient de la division de l'avant-dernier reste par le dernier.

Ligne	r	u	v
L_0	1071	1	0
L_1	462	0	1
$L_2 = L_0 - 2L_1$	147	1	-2
$L_3 = L_1 - 3L_2$	21	-3	7
$L_4 = L_2 - 7L_3$	0	22	-51

Corollaire E.10 **Conséquences du théorème de Bachet–Bézout**

Soient a et b deux entiers non tous les deux nuls, et $d = \operatorname{pgcd}(a, b)$.

(i) Tout diviseur commun de a et de b divise d.

(ii) Les entiers de la forme $au + bv$, avec $u, v \in \mathbb{Z}$, sont exactement les multiples de d.

(iii) Les entiers a et b sont premiers entre eux si et seulement s'il existe des entiers u et v tels que $au + bv = 1$.

Démonstration. (i) Écrivons $d = au + bv$. Un diviseur commun de a et de b divise cette combinaison, donc divise d.

(ii) Comme d divise a et b, il divise toute combinaison $au + bv$. Inversement, un multiple md de d s'écrit $md = a(mu) + b(mv)$.

(iii) Si $\operatorname{pgcd}(a, b) = 1$, le théorème donne directement u et v. Réciproquement, si $au + bv = 1$, alors d divise 1 par le point (ii), de sorte que $d = 1$.

□

Le point (iii) ne s'étend pas à d'autres valeurs que 1 : l'égalité $au + bv = 6$ montre seulement que $\operatorname{pgcd}(a, b)$ divise 6. Par exemple, $4 \times 3 + 2 \times (-3) = 6$, mais $\operatorname{pgcd}(4, 2) = 2$.

Remarque E.5 **Bachet ou Bézout ?**

Le résultat pour les entiers figure dans la deuxième édition (1624) des Problèmes plaisants et délectables qui se font par les nombres de Claude-Gaspard Bachet de Méziriac, un recueil d'énigmes arithmétiques. Étienne Bézout l'a étendu aux polynômes en 1779, dans un ouvrage sur les équations algébriques. L'usage, surtout en France, a longtemps retenu le seul nom de Bézout ; on parle de plus en plus du théorème de Bachet–Bézout, ou de l'identité de Bézout pour l'égalité $au + bv = d$ elle-même.

E.5 Le lemme de Gauss et l'unicité de la factorisation

Théorème E.11 **Lemme de Gauss**

Soient a, b et c des entiers, avec $a \neq 0$. Si $a \mid bc$ et si $\operatorname{pgcd}(a, b) = 1$, alors $a \mid c$.

Démonstration. Puisque a et b sont premiers entre eux, le théorème de Bachet–Bézout fournit des entiers u et v tels que $au + bv = 1$. En multipliant cette égalité par c, on obtient

$$c = acu + bcv.$$

L'entier a divise acu, évidemment, et il divise bcv puisqu'il divise bc par hypothèse. Il divise donc leur somme, qui est c. □

L'hypothèse $\operatorname{pgcd}(a, b) = 1$ est indispensable : 6 divise $4 \times 9 = 36$, mais 6 ne divise ni 4 ni 9. Le cas particulier où a est un nombre premier mérite d'être énoncé à part.

Corollaire E.12 **Lemme d'Euclide**

Soit p un nombre premier. Si $p \mid bc$, alors $p \mid b$ ou $p \mid c$. Plus généralement, si p divise un produit $a_1 a_2 \cdots a_k$, alors p divise au moins l'un des facteurs a_i.

Démonstration. Supposons que $p \nmid b$. Les seuls diviseurs positifs de p sont 1 et p ; comme p ne divise pas b, le seul diviseur commun positif de p et de b est 1, et $\operatorname{pgcd}(p, b) = 1$. Le lemme de Gauss donne alors $p \mid c$.

Pour un produit de k facteurs, on raisonne par récurrence sur k. Si $p \mid a_1(a_2 \cdots a_k)$, le premier cas montre que $p \mid a_1$ ou $p \mid a_2 \cdots a_k$; dans le second cas, l'hypothèse de récurrence s'applique au produit de $k - 1$ facteurs. □

Corollaire E.13 **Deux diviseurs premiers entre eux**

Soient a, b et c des entiers, avec a et b non nuls. Si $a \mid c$, $b \mid c$ et $\operatorname{pgcd}(a, b) = 1$, alors $ab \mid c$.

Démonstration. Écrivons $c = ak$. Alors $b \mid ak$ et $\operatorname{pgcd}(b, a) = 1$; le lemme de Gauss donne $b \mid k$, c'est-à-dire $k = bl$ pour un certain entier l. Ainsi $c = abl$, et $ab \mid c$. □

Là encore, l'hypothèse est nécessaire : $4 \mid 12$ et $6 \mid 12$, mais $24 \nmid 12$.

E.5.1 Le théorème fondamental de l'arithmétique

Théorème E.14 **Théorème fondamental de l'arithmétique**

Tout entier $n \geq 2$ s'écrit comme un produit de nombres premiers, et cette écriture est unique à l'ordre des facteurs près. Autrement dit, il existe des nombres premiers distincts $p_1 < p_2 < \cdots < p_k$ et des exposants $\alpha_1, \ldots, \alpha_k \geq 1$, déterminés de façon unique par n, tels que

$$n = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_k^{\alpha_k}.$$

Démonstration. L'existence a été établie à la proposition E.5. Montrons l'unicité, encore une fois à l'aide d'un plus petit contre-exemple. Supposons qu'il existe des entiers qui admettent deux factorisations différentes, et soit n le plus petit d'entre eux :

$$n = p_1 p_2 \cdots p_r = q_1 q_2 \cdots q_s,$$

où les p_i et les q_j sont des nombres premiers (pas forcément distincts), et où les deux listes ne sont pas formées des mêmes nombres, répétitions comprises.

Le nombre premier p_1 divise $n = q_1 q_2 \cdots q_s$. Par le lemme d'Euclide (corollaire E.12), il divise l'un des q_j. Or les seuls diviseurs de q_j supérieurs à 1 sont q_j lui-même ; donc $p_1 = q_j$. Quitte à renuméroter les q, on peut supposer que $p_1 = q_1$. En simplifiant par p_1, on obtient

$$\frac{n}{p_1} = p_2 \cdots p_r = q_2 \cdots q_s.$$

Si $n/p_1 = 1$, les deux produits sont vides ($r = s = 1$) et les factorisations de départ étaient identiques, ce qui est exclu. Sinon, n/p_1 est un entier compris entre 2 et $n - 1$ qui admet deux factorisations différentes (puisqu'on a retiré le même facteur des deux listes, qui étaient différentes). Cela contredit la minimalité de n. □

Remarque E.6 **Pourquoi 1 n'est pas premier**

Si l'on décrétait que 1 est premier, l'unicité tomberait aussitôt : $6 = 2 \times 3 = 1 \times 2 \times 3 = 1 \times 1 \times 2 \times 3$. Exclure 1 des nombres premiers n'est donc pas un caprice : c'est ce qui permet d'énoncer simplement le théorème fondamental. Il est aussi instructif de constater que l'unicité n'a rien d'évident. Dans l'ensemble des nombres pairs $\{2, 4, 6, 8, \ldots\}$, où l'on appellerait « premier » un nombre pair qui n'est pas produit de deux nombres pairs (2, 6, 10, 14, 18, …), on a $36 = 2 \times 18 = 6 \times 6$: deux factorisations différentes en « premiers ». C'est le lemme de Gauss, et à travers lui le théorème de Bachet–Bézout, qui empêche ce phénomène dans $\mathbb{Z}$.

La factorisation permet de lire directement le PGCD et le PPCM. Si l'on écrit a et b à l'aide des mêmes nombres premiers, en acceptant des exposants nuls,

$$a = p_1^{\alpha_1} \cdots p_k^{\alpha_k}, \qquad b = p_1^{\beta_1} \cdots p_k^{\beta_k},$$

un entier positif d divise a si et seulement si sa factorisation ne fait intervenir que des p_i, chacun avec un exposant au plus égal à α_i. (C'est une conséquence de l'unicité : si $a = dm$, la factorisation de a s'obtient en juxtaposant celles de d et de m.)

Proposition E.15 **PGCD et PPCM à partir des factorisations**

Avec les notations précédentes,

$$\operatorname{pgcd}(a,b) = \prod_{i=1}^{k} p_i^{\min(\alpha_i,\beta_i)}, \qquad \operatorname{ppcm}(a,b) = \prod_{i=1}^{k} p_i^{\max(\alpha_i,\beta_i)}.$$

En particulier, $\operatorname{pgcd}(a,b)\cdot\operatorname{ppcm}(a,b) = ab$.

Démonstration. Un diviseur commun de a et de b a, pour chaque p_i, un exposant au plus égal à α_i et au plus égal à β_i, donc au plus égal à $\min(\alpha_i,\beta_i)$. Le plus grand s'obtient en prenant exactement cet exposant pour chaque p_i. Le raisonnement est le même pour le PPCM, avec des exposants au moins égaux à α_i et à β_i. Enfin, pour deux nombres x et y, on a toujours $\min(x,y)+\max(x,y) = x+y$; en multipliant les deux formules, l'exposant de chaque p_i est donc $\alpha_i+\beta_i$, qui est son exposant dans ab. □

Exemple E.4 **Le PGCD et le PPCM de 252 et 198**

On a $252 = 2^2\cdot 3^2\cdot 7$ et $198 = 2\cdot 3^2\cdot 11$. En écrivant les deux nombres avec les mêmes facteurs premiers, $252 = 2^2\cdot 3^2\cdot 7^1\cdot 11^0$ et $198 = 2^1\cdot 3^2\cdot 7^0\cdot 11^1$, d'où

$$\operatorname{pgcd}(252,198) = 2^1\cdot 3^2 = 18, \qquad \operatorname{ppcm}(252,198) = 2^2\cdot 3^2\cdot 7\cdot 11 = 2772.$$

On vérifie que $18\times 2772 = 49\,896 = 252\times 198$.

E.5.2 Les équations diophantiennes linéaires

Une **équation diophantienne** est une équation dont on cherche les solutions entières. Le théorème de Bachet–Bézout et le lemme de Gauss permettent de résoudre complètement les plus simples d'entre elles.

Proposition E.16 **Équation ax + by = c**

Soient a et b des entiers non nuls, c un entier et $d = \operatorname{pgcd}(a,b)$. L'équation $ax+by = c$ a des solutions entières si et seulement si $d \mid c$. Dans ce cas, si (x_0,y_0) est une solution particulière, les solutions sont exactement les couples

$$x = x_0 + \frac{b}{d}\,t, \qquad y = y_0 - \frac{a}{d}\,t, \qquad t\in\mathbb{Z}\,.$$

Démonstration. Existence. Si $ax+by = c$, alors d, qui divise a et b, divise c. Réciproquement, si $c = dm$, le théorème de Bachet–Bézout fournit $au+bv = d$, et $(x_0,y_0) = (mu,mv)$ est une solution.

Forme des solutions. On vérifie directement que les couples proposés sont des solutions : $a(x_0+\frac{b}{d}t)+b(y_0-\frac{a}{d}t) = ax_0+by_0 = c$. Inversement, soit (x,y) une solution. En soustrayant $ax_0+by_0 = c$, on obtient $a(x-x_0) = -b(y-y_0)$, puis, en divisant par d,

$$a'(x-x_0) = -b'(y-y_0), \qquad \text{où } a' = \frac{a}{d} \text{ et } b' = \frac{b}{d}.$$

Les entiers a' et b' sont premiers entre eux (un diviseur commun $e>1$ de a' et b' ferait de de un diviseur commun de a et b plus grand que d). Puisque b' divise $a'(x-x_0)$ et que

$\operatorname{pgcd}(b', a') = 1$, le lemme de Gauss donne $b' \mid x - x_0 : x - x_0 = b't$ pour un certain entier t. En reportant, on obtient $a'b't = -b'(y - y_0)$, d'où $y - y_0 = -a't$. □

Exemple E.5 **Payer 100 dollars avec des jetons de 7 et de 5 dollars**

Dans un jeu, on dispose de jetons de 7 dollars et de 5 dollars. De combien de façons peut-on payer exactement 100 dollars ? On cherche les solutions positives ou nulles de $7x + 5y = 100$. Comme $\operatorname{pgcd}(7, 5) = 1$ divise 100, il y a des solutions entières. On remarque que $(x_0, y_0) = (0, 20)$ en est une ; à défaut, l'égalité de Bézout $7 \times (-2) + 5 \times 3 = 1$, multipliée par 100, aurait fourni $(-200, 300)$. Les solutions entières sont

$$x = 5t, \qquad y = 20 - 7t, \qquad t \in \mathbb{Z}\,.$$

Les conditions $x \geq 0$ et $y \geq 0$ imposent $0 \leq t \leq \frac{20}{7}$, soit $t \in \{0, 1, 2\}$. Il y a trois façons de payer : $(0, 20)$, $(5, 13)$ et $(10, 6)$.

E.6 Les congruences

Dans bien des situations, seul le reste d'une division nous intéresse. Quel jour de la semaine serons-nous dans 100 jours ? Quelle heure indiquera l'horloge dans 50 heures ? Un grand nombre est-il pair ? Les congruences, introduites par Gauss, fournissent une notation et des règles de calcul pour ces questions.

Définition E.4 **Congruence modulo n**

Soit $n \geq 1$ un entier. Deux entiers a et b sont **congrus modulo** n, ce qu'on note

$$a \equiv b \pmod{n},$$

si n divise $a - b$. L'entier n s'appelle le **module**.

Ainsi, $17 \equiv 2 \pmod 5$ puisque $5 \mid 15$, et $-3 \equiv 4 \pmod 7$ puisque $7 \mid -7$. Un entier est pair si et seulement s'il est congru à 0 modulo 2. Dire que $a \equiv 0 \pmod n$, c'est dire que $n \mid a$.

Proposition E.17 **Congruence et reste**

Soit $n \geq 1$. Deux entiers a et b sont congrus modulo n si et seulement s'ils ont le même reste dans la division euclidienne par n. En particulier, tout entier est congru modulo n à exactement un des entiers $0, 1, \ldots, n - 1$, à savoir son reste.

Démonstration. Écrivons $a = nq + r$ et $b = nq' + r'$, avec $0 \leq r, r' < n$. Alors $a - b = n(q - q') + (r - r')$. Si $r = r'$, on voit que $n \mid a - b$. Réciproquement, si $n \mid a - b$, alors n divise $(a - b) - n(q - q') = r - r'$. Or $-n < r - r' < n$, et le seul multiple de n dans cet intervalle est 0 ; donc $r = r'$. Pour la dernière affirmation, $a = nq + r$ donne $a - r = nq$, donc $a \equiv r \pmod n$; et deux entiers distincts de $\{0, \ldots, n - 1\}$ ne sont pas congrus, puisqu'ils sont eux-mêmes leurs propres restes. □

Les entiers se répartissent donc en n familles, ou **classes de congruence**, selon leur reste.

Modulo 7, par exemple, la classe de 3 est $\{\dots, -11, -4, 3, 10, 17, \dots\}$. Calculer modulo n, c'est calculer en ne gardant que la classe des nombres, comme on le fait sur une horloge, où 13 h est « la même chose » que 1 h (modulo 12).

Proposition E.18 **Règles de calcul**

Soient $n \geq 1$ et a, b, c, d des entiers.

(i) La congruence modulo n est une relation d'équivalence : $a \equiv a$; si $a \equiv b$, alors $b \equiv a$; si $a \equiv b$ et $b \equiv c$, alors $a \equiv c$ (modulo n).

(ii) Si $a \equiv b \pmod{n}$ et $c \equiv d \pmod{n}$, alors

$$a + c \equiv b + d, \qquad a - c \equiv b - d, \qquad ac \equiv bd \pmod{n}.$$

(iii) Si $a \equiv b \pmod{n}$, alors $a^k \equiv b^k \pmod{n}$ pour tout entier $k \geq 1$.

Démonstration. Le point (i) et les deux premières congruences du point (ii) découlent directement de la proposition E.1 ; leur démonstration fait l'objet de l'exercice E.9. Démontrons la règle du produit. Par hypothèse, $a - b = nk$ et $c - d = nj$ pour des entiers k et j. L'astuce consiste à ajouter et retrancher bc :

$$ac - bd = ac - bc + bc - bd = (a - b)\,c + b\,(c - d) = n\,(kc + bj).$$

Ainsi $n \mid ac - bd$, c'est-à-dire $ac \equiv bd \pmod{n}$.

Le point (iii) s'obtient en appliquant la règle du produit $k - 1$ fois : de $a \equiv b$ et $a \equiv b$, on tire $a^2 \equiv b^2$; de $a^2 \equiv b^2$ et $a \equiv b$, on tire $a^3 \equiv b^3$; et ainsi de suite. □

Ces règles autorisent, dans une somme, un produit ou une puissance, à remplacer n'importe quel nombre par un autre de la même classe, en général plus petit. C'est ce qui rend les congruences si efficaces.

Exemple E.6 **Quelques calculs modulaires**

Nous sommes un jeudi. Comme $100 = 14 \times 7 + 2$, on a $100 \equiv 2 \pmod{7}$, et dans 100 jours nous serons deux jours après un jeudi, soit un samedi.
Le dernier chiffre de 7^{100} est son reste modulo 10. Or $7^2 = 49 \equiv -1 \pmod{10}$; donc $7^4 \equiv (-1)^2 = 1$, et $7^{100} = (7^4)^{25} \equiv 1^{25} = 1 \pmod{10}$. Le nombre 7^{100}, qui compte 85 chiffres, se termine par 1.
Les critères de divisibilité s'expliquent de la même façon. Comme $10 \equiv 1 \pmod{9}$, on a $10^k \equiv 1 \pmod{9}$ pour tout k, et un nombre $\overline{c_m \cdots c_1 c_0} = c_m 10^m + \cdots + c_1 10 + c_0$ est congru modulo 9 à la somme $c_m + \cdots + c_1 + c_0$ de ses chiffres. Ainsi, 738 261 est congru à $7 + 3 + 8 + 2 + 6 + 1 = 27 \equiv 0 \pmod{9}$: il est divisible par 9. Le même raisonnement, modulo 3, justifie le critère de divisibilité par 3.

Pour la classe E.2 **Un tour de magie : deviner le chiffre biffé**

Ce tour, qui repose entièrement sur le critère de divisibilité par 9, captive les élèves et leur donne une bonne raison de chercher à comprendre une propriété des congruences.

Le déroulement. L'enseignant tourne le dos à la classe. Quelques élèves choisissent chacun un entier différent de 1, sans le dire à voix haute, et la classe calcule le

produit de tous ces nombres à la calculatrice. L'enseignant propose ensuite, sous un prétexte quelconque, de multiplier encore le résultat par un ou deux nombres : « multiplions par 18, car nous sommes 18 dans la classe », ou « multiplions par 3 et par 6, car il y a 3 filles et 6 garçons dans la première rangée ». Les élèves biffent alors tous les zéros du résultat, puis un autre chiffre de leur choix, et montrent à l'enseignant les chiffres qui restent, dans l'ordre qu'ils veulent. L'enseignant annonce aussitôt le chiffre biffé.

Le secret. Le facteur ajouté par l'enseignant contient 9 ($18 = 2 \times 9$ et $3 \times 6 = 18$). Le produit P est donc un multiple de 9, et la somme de ses chiffres est congrue à 0 modulo 9 (exemple E.6). Biffer les zéros ou mélanger les chiffres ne change pas cette somme. Si S est la somme des chiffres montrés et c le chiffre biffé, on a donc $S + c \equiv 0 \pmod 9$. Comme c est compris entre 1 et 9 (les zéros ont déjà disparu), une seule valeur convient : c est le nombre qui manque à S pour atteindre le multiple de 9 suivant, et $c = 9$ lorsque S est lui-même un multiple de 9.

Un exemple. Les élèves choisissent 7, 12, 25 et 13, dont le produit vaut 27 300 ; multiplié par 18, il donne 491 400. Une fois les zéros biffés, il reste 4914. Un élève biffe le 1 et montre 4, 9, 4. L'enseignant calcule $4 + 9 + 4 = 17$; le multiple de 9 suivant est 18, donc le chiffre biffé est 1. Si l'élève avait biffé le 9, il aurait montré 4, 1, 4, de somme 9, et l'enseignant aurait répondu 9.

Un tour qui fonctionne presque sans aide. Le facteur ajouté par l'enseignant est une assurance, mais il est rarement nécessaire. Le produit est un multiple de 9 dès que deux des nombres choisis sont des multiples de 3, ou que l'un d'eux est un multiple de 9. Chaque nombre a environ une chance sur trois d'être un multiple de 3 ; avec une dizaine de nombres choisis au hasard, le produit est un multiple de 9 dans plus de 90 % des cas. Pour un premier essai, on peut donc tenter le tour sans rien ajouter, et garder le facteur 18 pour les fois où il faut une réussite assurée. Une précaution pratique s'impose aussi : il faut limiter le nombre et la taille des facteurs, pour que le produit s'affiche en entier sur la calculatrice, sans passer en notation scientifique.

Après le tour. On peut laisser les élèves chercher le secret, puis le leur faire refaire à tour de rôle. Le tour amène naturellement plusieurs questions. Pourquoi faut-il interdire de biffer un zéro ? (On ne pourrait pas distinguer un 0 d'un 9 biffé.) Fonctionnerait-il avec la somme des chiffres modulo 3 ? (Non : un chiffre biffé entre 1 et 9 ne serait connu qu'à un multiple de 3 près.) Pourquoi le nombre 18 convient-il, et pas 12 ?

Remarque E.7 **On ne peut pas toujours simplifier**

L'addition, la soustraction et la multiplication se comportent bien, mais pas la division. On a $2 \times 3 \equiv 2 \times 8 \pmod{10}$, puisque $6 \equiv 16$, et pourtant $3 \not\equiv 8 \pmod{10}$. La simplification par a est permise lorsque $\operatorname{pgcd}(a, n) = 1$: si $ax \equiv ay \pmod n$, alors $n \mid a(x - y)$, et le lemme de Gauss donne $n \mid x - y$. Lorsque n est premier, on peut donc simplifier par tout a non divisible par n.

E.6.1 Inverses modulaires et congruences linéaires

Définition E.5 **Inverse modulo n**

Soit $n \geq 2$. Un entier a' est un **inverse de a modulo n** si $aa' \equiv 1 \pmod{n}$. On dit alors que a est **inversible modulo n**.

Par exemple, $3 \times 5 = 15 \equiv 1 \pmod{7}$: 5 est un inverse de 3 modulo 7. En revanche, 2 n'a pas d'inverse modulo 4, car $2a'$ est toujours pair et ne peut pas être congru à 1 modulo 4.

Proposition E.19 **Quand un entier est-il inversible ?**

Soit $n \geq 2$. Un entier a est inversible modulo n si et seulement si $\gcd(a, n) = 1$. Dans ce cas, son inverse est unique modulo n, et on l'obtient à l'aide de l'algorithme d'Euclide étendu.

Démonstration. Dire que $aa' \equiv 1 \pmod{n}$, c'est dire qu'il existe un entier k tel que $aa' - 1 = kn$, c'est-à-dire $aa' + n(-k) = 1$. Par le point (iii) du corollaire E.10, de tels entiers a' et k existent si et seulement si $\gcd(a, n) = 1$. Les coefficients de Bézout u et v de l'égalité $au + nv = 1$ fournissent l'inverse $a' = u$.

Si a' et a'' sont deux inverses de a, alors $a' \equiv a'(aa'') = (a'a)a'' \equiv a'' \pmod{n}$. □

Exemple E.7 **L'inverse de 7 modulo 26**

On applique l'algorithme d'Euclide à 26 et 7 : $26 = 3 \times 7 + 5$, $7 = 1 \times 5 + 2$, $5 = 2 \times 2 + 1$. En remontant,

$$\begin{aligned} 1 &= 5 - 2 \times 2 = 5 - 2\,(7 - 5) = 3 \times 5 - 2 \times 7 \\ &= 3\,(26 - 3 \times 7) - 2 \times 7 = 3 \times 26 - 11 \times 7. \end{aligned}$$

Ainsi $7 \times (-11) \equiv 1 \pmod{26}$, et l'inverse de 7 est $-11 \equiv 15 \pmod{26}$. On vérifie : $7 \times 15 = 105 = 4 \times 26 + 1$. Ce calcul est celui qu'il faut faire pour déchiffrer un message codé par le chiffrement affine $x \mapsto 7x + 3 \pmod{26}$, où chaque lettre est remplacée par son rang dans l'alphabet.

Proposition E.20 **Congruences linéaires**

Soient $n \geq 2$, a et b des entiers, et $d = \gcd(a, n)$. La congruence $ax \equiv b \pmod{n}$ a des solutions si et seulement si $d \mid b$. Dans ce cas, ses solutions forment une seule classe modulo n/d, c'est-à-dire d classes distinctes modulo n.

Démonstration. La congruence $ax \equiv b \pmod{n}$ équivaut à l'existence d'un entier y tel que $ax - b = ny$, c'est-à-dire $ax + n(-y) = b$. C'est une équation diophantienne linéaire ; par la proposition E.16, elle a des solutions si et seulement si $d \mid b$, et ses solutions en x sont alors les $x = x_0 + \frac{n}{d}\,t$, $t \in \mathbb{Z}$. Ce sont exactement les entiers congrus à x_0 modulo n/d. Parmi $0, 1, \ldots, n-1$, on en trouve d : $x_0, x_0 + \frac{n}{d}, \ldots, x_0 + (d-1)\frac{n}{d}$ (réduits modulo n). □

Lorsque $\gcd(a, n) = 1$, la solution est unique modulo n : il suffit de multiplier les deux membres par l'inverse a' de a, ce qui donne $x \equiv a'b \pmod{n}$.

Exemple E.8 **Une congruence à deux solutions**

Ici $d = \text{pgcd}(4, 10) = 2$ divise 6, et il y a deux solutions modulo 10. En divisant tout par 2 (y compris le module), la congruence devient $2x \equiv 3 \pmod 5$. L'inverse de 2 modulo 5 est 3 ($2 \times 3 = 6 \equiv 1$), d'où $x \equiv 3 \times 3 = 9 \equiv 4 \pmod 5$. Les solutions modulo 10 sont donc $x \equiv 4$ et $x \equiv 9 \pmod{10}$. On vérifie que $4 \times 4 = 16 \equiv 6$ et $4 \times 9 = 36 \equiv 6 \pmod{10}$. Par contre, la congruence $4x \equiv 5 \pmod{10}$ n'a pas de solution, puisque $2 \nmid 5$: le membre de gauche est toujours pair modulo 10, le membre de droite impair.

Note historique E.3 **Carl Friedrich Gauss**



Carl Friedrich Gauss (1777–1855), portrait de Christian Albrecht Jensen (1840). Domaine public, Wikimedia Commons.

Carl Friedrich Gauss naît en 1777 à Brunswick, dans une famille modeste. Selon une anecdote célèbre, sans doute embellie, son instituteur aurait demandé à la classe d'additionner les entiers de 1 à 100, et le jeune Gauss aurait aussitôt donné la réponse, 5050, en regroupant les termes deux par deux ($1 + 100$, $2 + 99$, …). Repéré pour son talent, il étudie grâce au soutien du duc de Brunswick. À dix-neuf ans, il montre qu'on peut construire à la règle et au compas le polygone régulier à 17 côtés, un problème resté ouvert depuis les Grecs. Cette découverte, raconte-t-on, le décide à se consacrer aux mathématiques plutôt qu'aux langues anciennes.

En 1801, à vingt-quatre ans, il publie les Disquisitiones arithmeticae (« Recherches arithmétiques ») [Gau01], l'ouvrage qui fonde la théorie des nombres moderne. Gauss y introduit la notation $a \equiv b \pmod n$ et le calcul des congruences, dont il fait un langage systématique. Il y énonce et démontre clairement l'unicité de la factorisation en nombres premiers, que les Éléments d'Euclide ne formulaient pas sous cette forme. Il y donne aussi la première démonstration de la loi de réciprocité quadratique, qu'il appelait le « théorème d'or ». C'est également dans cet ouvrage qu'apparaît la notation $\varphi(n)$ pour l'indicatrice d'Euler (section E.8). Le lemme qui porte son nom dans la tradition française (théorème E.11) figure en germe chez Euclide, mais c'est Gauss qui en fait l'outil central de l'arithmétique.

L'œuvre de Gauss déborde largement la théorie des nombres : astronomie (en 1801, il calcule l'orbite de l'astéroïde Cérès, perdu de vue après quelques semaines d'observation, ce qui permet de le retrouver ; il s'appuie pour cela sur la méthode des moindres carrés, qu'il avait mise au point), géodésie, magnétisme, géométrie différentielle, statistique (la loi normale porte aussi le nom de loi de Gauss) et, nous l'avons vu à l'annexe D, résolution des systèmes linéaires. Surnommé de son vivant le « prince des mathématiciens », il aurait affirmé que les mathématiques sont la reine des sciences, et l'arithmétique la reine des mathématiques [ORf].

E.7 Le petit théorème de Fermat

Calculons les puissances successives de 3 modulo 7 :

$$3^1 \equiv 3, \quad 3^2 \equiv 2, \quad 3^3 \equiv 6, \quad 3^4 \equiv 4, \quad 3^5 \equiv 5, \quad 3^6 \equiv 1 \pmod 7.$$

On retombe sur 1 à l'exposant $6 = 7 - 1$. Si l'on recommence avec 2, on trouve $2^3 \equiv 1$, et donc $2^6 \equiv 1$ aussi. Ce n'est pas une coïncidence.

Théorème E.21 **Petit théorème de Fermat**

Soient p un nombre premier et a un entier non divisible par p. Alors

$$a^{p-1} \equiv 1 \pmod p.$$

Avant de démontrer ce théorème, observons ce qui se passe pour $p = 7$ et $a = 3$. Multiplions par 3 chacun des nombres $1, 2, 3, 4, 5, 6$, et réduisons modulo 7 :

$$3,\ 6,\ 9 \equiv 2,\ 12 \equiv 5,\ 15 \equiv 1,\ 18 \equiv 4.$$

On retrouve les mêmes six nombres, dans un autre ordre. Le produit des six nombres de gauche est donc congru au produit des six nombres de départ :

$$(3 \cdot 1)(3 \cdot 2)(3 \cdot 3)(3 \cdot 4)(3 \cdot 5)(3 \cdot 6) \equiv 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \pmod 7,$$

c'est-à-dire $3^6 \cdot 6! \equiv 6! \pmod 7$. Comme $6! = 720$ n'est pas divisible par 7, on peut simplifier (remarque E.7), et il reste $3^6 \equiv 1 \pmod 7$. La démonstration générale suit exactement ce chemin.

Démonstration. Considérons les $p - 1$ entiers

$$a,\ 2a,\ 3a,\ \ldots,\ (p-1)a.$$

Aucun d'eux n'est divisible par p. En effet, si $p \mid ka$ avec $1 \leq k \leq p - 1$, le lemme d'Euclide donnerait $p \mid k$ ou $p \mid a$; or $p \nmid a$ par hypothèse, et $p \nmid k$ puisque $0 < k < p$. Leurs restes modulo p sont donc tous dans $\{1, 2, \ldots, p-1\}$.

Leurs restes sont deux à deux distincts. Supposons en effet que $ia \equiv ja \pmod p$ avec $1 \leq i < j \leq p - 1$. Alors $p \mid (j - i)a$; comme $p \nmid a$, le lemme d'Euclide donne $p \mid j - i$, ce qui est impossible puisque $0 < j - i < p$.

Ainsi, les restes de $a, 2a, \ldots, (p-1)a$ sont $p - 1$ nombres distincts de l'ensemble $\{1, 2, \ldots, p-1\}$, qui compte exactement $p - 1$ éléments : ce sont donc tous les éléments de cet ensemble, dans un certain ordre. En multipliant ces congruences entre elles (point (ii) de la proposition E.18), on obtient

$$a \cdot 2a \cdot 3a \cdots (p-1)a \equiv 1 \cdot 2 \cdot 3 \cdots (p-1) \pmod p, \qquad \text{soit} \qquad a^{p-1}\,(p-1)! \equiv (p-1)! \pmod p.$$

Enfin, p ne divise pas $(p-1)! = 1 \cdot 2 \cdots (p-1)$, encore par le lemme d'Euclide, puisqu'il ne divise aucun des facteurs. Donc $\operatorname{pgcd}\big((p-1)!, p\big) = 1$, et l'on peut simplifier par $(p-1)!$ (remarque E.7) : $a^{p-1} \equiv 1 \pmod p$. □

Corollaire E.22 **Forme générale**

Si p est premier, alors $a^p \equiv a \pmod{p}$ pour tout entier a.

Démonstration. Si $p \nmid a$, on multiplie par a la congruence $a^{p-1} \equiv 1$. Si $p \mid a$, les deux membres sont congrus à 0. □

Exemple E.9 **Réduire un exposant**

Cherchons le reste de 2^{100} dans la division par 13. Comme 13 est premier et ne divise pas 2, on a $2^{12} \equiv 1 \pmod{13}$. Or $100 = 8 \times 12 + 4$, d'où

$$2^{100} = (2^{12})^8 \cdot 2^4 \equiv 1^8 \cdot 16 \equiv 3 \pmod{13}.$$

De façon générale, modulo un nombre premier p, on peut réduire l'exposant modulo $p - 1$ (lorsque la base n'est pas un multiple de p).

Remarque E.8 **Un test de primalité, et ses limites**

Le petit théorème de Fermat fournit un moyen de démontrer qu'un nombre n'est pas premier, sans le factoriser : si $2^{n-1} \not\equiv 1 \pmod{n}$, alors n est composé. Par exemple, $2^{14} = 16\,384 \equiv 4 \pmod{15}$, ce qui prouve que 15 n'est pas premier. La réciproque est fausse : $341 = 11 \times 31$ vérifie $2^{340} \equiv 1 \pmod{341}$ (exercice E.14). Des versions plus fines de ce test, comme celui de Miller et Rabin, sont utilisées par les ordinateurs pour produire les très grands nombres premiers dont a besoin la cryptographie.

Pierre de Fermat, magistrat à Toulouse et mathématicien amateur, énonce ce résultat en 1640 dans une lettre à Bernard Frénicle de Bessy, sans démonstration, en ajoutant qu'il lui enverrait la preuve s'il ne craignait d'être trop long. La première démonstration publiée est due à Leonhard Euler, en 1736. Euler le généralisera à un module quelconque, comme nous allons le voir.

E.8 L'indicatrice d'Euler et le théorème d'Euler

Lorsque le module n n'est pas premier, le petit théorème de Fermat tombe en défaut : $2^3 = 8 \equiv 0 \pmod{4}$. Pour le généraliser, il faut compter les entiers qui « se comportent bien » modulo n, c'est-à- dire ceux qui sont inversibles.

Définition E.6 **Indicatrice d'Euler**

Pour un entier $n \geq 1$, on note $\varphi(n)$ le nombre d'entiers k compris entre 1 et n qui sont premiers avec n. La fonction φ s'appelle l'**indicatrice d'Euler**.

Par exemple, les entiers de 1 à 12 premiers avec 12 sont 1, 5, 7 et 11 ; donc $\varphi(12) = 4$. Les premières valeurs sont données dans le tableau E.2.

Table E.2 : Les premières valeurs de l'indicatrice d'Euler.

n	1	2	3	4	5	6	7	8	9	10	11	12
$\varphi(n)$	1	1	2	2	4	2	6	4	6	4	10	4

Proposition E.23 **Calcul de l'indicatrice d'Euler**

(i) Si p est premier, $\varphi(p) = p - 1$.

(ii) Si p est premier et $k \geq 1$, $\varphi(p^k) = p^k - p^{k-1}$.

(iii) Si $\operatorname{pgcd}(m, n) = 1$, alors $\varphi(mn) = \varphi(m)\,\varphi(n)$.

(iv) Par conséquent, si $n = p_1^{\alpha_1} \cdots p_k^{\alpha_k}$ est la factorisation de n,

$$\varphi(n) = n\left(1 - \frac{1}{p_1}\right) \cdots \left(1 - \frac{1}{p_k}\right).$$

Démonstration. (i) Les entiers $1, 2, \ldots, p-1$ ne sont pas divisibles par p ; comme les seuls diviseurs positifs de p sont 1 et p, ils sont tous premiers avec p. Seul p lui-même ne l'est pas.

(ii) La démonstration fait l'objet de l'exercice E.16.

(iii) Ce point est une conséquence du théorème des restes chinois ; il est démontré au corollaire E.26.

(iv) Par (iii), appliqué de proche en proche aux facteurs $p_i^{\alpha_i}$, qui sont premiers entre eux deux à deux, puis par (ii),

$$\varphi(n) = \prod_{i=1}^{k} \left(p_i^{\alpha_i} - p_i^{\alpha_i - 1}\right) = \prod_{i=1}^{k} p_i^{\alpha_i} \left(1 - \frac{1}{p_i}\right) = n \prod_{i=1}^{k} \left(1 - \frac{1}{p_i}\right).$$

□

Par exemple, $\varphi(100) = 100\left(1 - \frac{1}{2}\right)\left(1 - \frac{1}{5}\right) = 40$: parmi les entiers de 1 à 100, 40 ne sont divisibles ni par 2 ni par 5.

Théorème E.24 **Théorème d'Euler**

Soient $n \geq 1$ un entier et a un entier premier avec n. Alors

$$a^{\varphi(n)} \equiv 1 \pmod{n}.$$

Démonstration. La démonstration reprend celle du petit théorème de Fermat, en remplaçant la liste $1, 2, \ldots, p-1$ par la liste des entiers premiers avec n. Notons $k_1, k_2, \ldots, k_m$ les entiers de $\{1, \ldots, n\}$ premiers avec n, où $m = \varphi(n)$, et considérons les produits

$$ak_1,\ ak_2,\ \ldots,\ ak_m.$$

Chacun est premier avec n. Si un nombre premier p divisait à la fois n et ak_i, il diviserait a ou k_i (lemme d'Euclide), et serait donc un diviseur commun de n et de a, ou de n et de k_i ; c'est impossible, puisque ces paires sont premières entre elles. Le reste de ak_i modulo n est donc lui aussi premier avec n (un diviseur commun de n et du reste diviserait aussi ak_i). Il fait donc partie de la liste $k_1, \ldots, k_m$.

Leurs restes sont deux à deux distincts. Si $ak_i \equiv ak_j \pmod{n}$, on peut simplifier par a, puisque $\operatorname{pgcd}(a, n) = 1$ (remarque E.7) ; on obtient $k_i \equiv k_j \pmod{n}$, d'où $k_i = k_j$, puisque ces deux nombres sont compris entre 1 et n.

Les restes de $ak_1, \dots, ak_m$ sont donc les nombres $k_1, \dots, k_m$, dans un certain ordre. En multipliant,

$$a^m\, k_1 k_2 \cdots k_m \equiv k_1 k_2 \cdots k_m \pmod{n}.$$

Le produit $k_1 \cdots k_m$ est premier avec n, puisque chacun de ses facteurs l'est ; on peut donc simplifier, et $a^m \equiv 1 \pmod{n}$. □

Lorsque $n = p$ est premier, $\varphi(p) = p - 1$, et l'on retrouve le petit théorème de Fermat. Le théorème d'Euler permet, comme celui-ci, de réduire les exposants : si $\operatorname{pgcd}(a, n) = 1$, on peut remplacer l'exposant par son reste modulo $\varphi(n)$.

Exemple E.10 **Le dernier chiffre d'une grande puissance**

Comme $\operatorname{pgcd}(3, 10) = 1$ et $\varphi(10) = 4$, on a $3^4 \equiv 1 \pmod{10}$ (en effet, $3^4 = 81$). Puisque $2026 = 4 \times 506 + 2$,

$$3^{2026} = \left(3^4\right)^{506} \cdot 3^2 \equiv 1 \cdot 9 = 9 \pmod{10}.$$

Le nombre 3^{2026} se termine par 9. Pour obtenir les deux derniers chiffres d'une puissance, on travaillera de même modulo 100, avec $\varphi(100) = 40$.

Remarque E.9 **Une application : le chiffrement RSA**

Le chiffrement RSA, publié en 1977 par Ronald Rivest, Adi Shamir et Leonard Adleman, est une application directe du théorème d'Euler. Le destinataire choisit deux grands nombres premiers p et q (de plusieurs centaines de chiffres) et calcule $n = pq$ et $\varphi(n) = (p-1)(q-1)$. Il choisit un exposant e premier avec $\varphi(n)$ et calcule son inverse d modulo $\varphi(n)$ par l'algorithme d'Euclide étendu. Il publie la clé (n, e) et garde d secret. Pour envoyer un message représenté par un entier $m < n$, on transmet $c \equiv m^e \pmod{n}$. Le destinataire calcule alors $c^d \equiv m^{ed} \pmod{n}$. Comme $ed = 1 + k\,\varphi(n)$ pour un certain entier k, le théorème d'Euler donne, si $\operatorname{pgcd}(m, n) = 1$,

$$m^{ed} = m \cdot \left(m^{\varphi(n)}\right)^k \equiv m \cdot 1^k = m \pmod{n},$$

et le message est retrouvé. (On montre que le résultat reste vrai lorsque m n'est pas premier avec n, à l'aide du théorème des restes chinois.)
Avec des nombres minuscules : $p = 3$, $q = 11$, $n = 33$, $\varphi(n) = 20$, $e = 3$ et $d = 7$ (car $3 \times 7 = 21 \equiv 1 \pmod{20}$). Le message $m = 4$ devient $c \equiv 4^3 = 64 \equiv 31 \pmod{33}$. Pour déchiffrer, on calcule $31^7 \pmod{33}$; comme $31 \equiv -2$, cela donne $(-2)^7 = -128 \equiv -128 + 4 \times 33 = 4 \pmod{33}$. La sécurité repose sur le fait que, pour calculer d, il faut connaître $\varphi(n)$, donc factoriser n, ce qu'on ne sait pas faire en un temps raisonnable lorsque p et q sont très grands (remarque E.4).

E.9 Le théorème des restes chinois

Un problème du Sunzi suanjing, un manuel chinois de calcul rédigé entre le IIIe siècle et le Ve siècle, demande : « Soit des objets dont on ignore le nombre. Comptés par trois, il en reste deux ; comptés par cinq, il en reste trois ; comptés par sept, il en reste deux. Combien y a-t-il d'objets ? » En langage de congruences, on cherche x tel que

$$x \equiv 2 \pmod 3, \qquad x \equiv 3 \pmod 5, \qquad x \equiv 2 \pmod 7.$$

Théorème E.25 **Théorème des restes chinois**

Soient m et n deux entiers supérieurs ou égaux à 1, premiers entre eux. Pour tous les entiers a et b, le système

$$x \equiv a \pmod m, \qquad x \equiv b \pmod n$$

a des solutions, et ces solutions forment une seule classe modulo mn : si x_0 est une solution, les solutions sont exactement les entiers $x \equiv x_0 \pmod{mn}$.
Plus généralement, si $n_1, n_2, \dots, n_k$ sont premiers entre eux deux à deux, le système $x \equiv a_i \pmod{n_i}$ $(i = 1, \dots, k)$ a des solutions, qui forment une seule classe modulo $n_1 n_2 \cdots n_k$.

Démonstration. Existence. Puisque $\operatorname{pgcd}(m, n) = 1$, le théorème de Bachet–Bézout fournit des entiers u et v tels que $mu + nv = 1$. Cette égalité dit deux choses :

$$nv = 1 - mu \equiv 1 \pmod m \qquad \text{et} \qquad mu = 1 - nv \equiv 1 \pmod n.$$

Le nombre nv « vaut 1 modulo m et 0 modulo n », et le nombre mu « vaut 0 modulo m et 1 modulo n ». Posons alors

$$x_0 = a \cdot nv + b \cdot mu.$$

Modulo m, on a $nv \equiv 1$ et $mu \equiv 0$, d'où $x_0 \equiv a \cdot 1 + b \cdot 0 = a$. Modulo n, de même, $x_0 \equiv a \cdot 0 + b \cdot 1 = b$. Ainsi x_0 est une solution.

Unicité modulo mn. Si $x \equiv x_0 \pmod{mn}$, alors $mn \mid x - x_0$, donc $m \mid x - x_0$ et $n \mid x - x_0$: x est aussi une solution. Réciproquement, si x est une solution, alors $x \equiv a \equiv x_0 \pmod m$ et $x \equiv b \equiv x_0 \pmod n$. Les entiers m et n divisent donc tous deux $x - x_0$; comme ils sont premiers entre eux, le corollaire E.13 donne $mn \mid x - x_0$.

Cas de k modules. On procède par récurrence sur k. Les deux premières congruences équivalent, par ce qui précède, à une seule congruence $x \equiv c \pmod{n_1 n_2}$. Le nombre $n_1 n_2$ est premier avec n_3 : un nombre premier qui diviserait $n_1 n_2$ et n_3 diviserait n_1 ou n_2 (lemme d'Euclide), alors que ceux-ci sont premiers avec n_3. On se ramène ainsi à un système de $k - 1$ congruences dont les modules $n_1 n_2, n_3, \dots, n_k$ sont encore premiers entre eux deux à deux. □

Exemple E.11 **Le problème de Sunzi**

Résolvons d'abord les deux premières congruences. La condition $x \equiv 3 \pmod 5$ s'écrit $x = 3 + 5k$. En reportant dans la première, $3 + 5k \equiv 2 \pmod 3$, soit $2k \equiv -1 \equiv 2 \pmod 3$, d'où $k \equiv 1 \pmod 3$ (on simplifie par 2, qui est premier avec 3). Avec $k = 1 + 3j$, on obtient $x = 8 + 15j$, c'est-à-dire $x \equiv 8 \pmod{15}$. On reporte dans la troisième : $8 + 15j \equiv 2 \pmod 7$. Comme $8 \equiv 1$ et $15 \equiv 1 \pmod 7$, cela donne $1 + j \equiv 2$, soit $j \equiv 1 \pmod 7$. Avec $j = 1 + 7i$, on obtient $x = 23 + 105i$. Les solutions sont les $x \equiv 23 \pmod{105}$; le plus petit nombre d'objets possible est 23. On vérifie que $23 = 7 \times 3 + 2 = 4 \times 5 + 3 = 3 \times 7 + 2$. On peut aussi utiliser la formule de la démonstration. Pour les deux premières congruences, avec $m = 3$ et $n = 5$, l'égalité $3 \times 2 + 5 \times (-1) = 1$ donne $u = 2$, $v = -1$, et $x_0 = 2 \times 5 \times (-1) + 3 \times 3 \times 2 = -10 + 18 = 8$.

Le théorème dit, en somme, que la donnée du reste modulo m et du reste modulo n détermine le reste modulo mn, et que toutes les paires de restes sont possibles. Le tableau E.3 le montre pour $m = 3$ et $n = 5$: chacun des 15 entiers $0, 1, \ldots, 14$ occupe une case différente, et toutes les cases sont remplies.

Table E.3 : Les entiers de 0 à 14, rangés selon leur reste modulo 3 (lignes) et leur reste modulo 5 (colonnes). Chaque case contient exactement un entier : c'est le théorème des restes chinois pour $m = 3$ et $n = 5$.

	$\equiv 0 \pmod 5$	$\equiv 1$	$\equiv 2$	$\equiv 3$	$\equiv 4$
$\equiv 0 \pmod 3$	0	6	12	3	9
$\equiv 1$	10	1	7	13	4
$\equiv 2$	5	11	2	8	14

On peut lire ce tableau en suivant les entiers dans l'ordre : on avance d'une ligne et d'une colonne à chaque pas, en revenant au début quand on sort du tableau, comme sur deux horloges qui tournent en même temps. Si les modules n'étaient pas premiers entre eux, ce parcours ne passerait pas par toutes les cases : modulo 2 et 4, par exemple, un entier pair n'est jamais congru à 1 modulo 4, et le système $x \equiv 0 \pmod 2$, $x \equiv 1 \pmod 4$ n'a aucune solution.

Corollaire E.26 **L'indicatrice d'Euler est multiplicative**

Si $\text{pgcd}(m, n) = 1$, alors $\varphi(mn) = \varphi(m)\,\varphi(n)$.

Démonstration. Par le théorème des restes chinois, l'application qui associe à chaque entier $x \in \{0, 1, \ldots, mn - 1\}$ le couple formé de son reste modulo m et de son reste modulo n est une bijection vers l'ensemble des couples (r, s), avec $0 \leq r < m$ et $0 \leq s < n$ (comme dans le tableau E.3). Or x est premier avec mn si et seulement s'il est premier avec m et avec n (un nombre premier qui divise mn divise m ou n). Par ailleurs, x est premier avec m si et seulement si son reste r l'est, puisque x et r ont les mêmes diviseurs communs avec m (lemme E.6) ; de même pour n. La bijection fait donc correspondre les entiers premiers avec mn aux couples (r, s) où r est premier avec m et s premier avec n. Il y a $\varphi(m)$ choix pour r et $\varphi(n)$ choix pour s, donc $\varphi(m)\varphi(n)$ couples. (On compte ici les entiers de 0 à $mn - 1$ plutôt que de 1 à mn, ce qui ne change rien : 0 et mn ne sont premiers avec mn que si $mn = 1$, cas évident.) □

Note historique E.4 Du Sunzi suanjing aux ordinateurs

Le problème de Sunzi n'est accompagné que d'une méthode propre aux modules 3, 5 et 7. La méthode générale, appelée dayan qiuyi shu (« méthode de la grande extension pour trouver l'unité »), est exposée en 1247 par Qin Jiushao dans ses Neuf sections de mathématiques ; elle revient à calculer des inverses modulaires par un procédé équivalent à l'algorithme d'Euclide étendu. En Europe, le résultat apparaît chez Fibonacci (1202), puis est démontré en toute généralité par Euler et par Gauss, qui lui consacre plusieurs articles des Disquisitiones. Le théorème sert aujourd'hui à accélérer les calculs sur de grands nombres : au lieu de calculer modulo $n = pq$, on calcule séparément modulo p et modulo q, avec des nombres deux fois plus courts, puis on recolle les résultats. C'est ainsi que les logiciels effectuent le déchiffrement RSA, environ quatre fois plus vite qu'en travaillant directement modulo n.

Exercices

Exercice E.1. Soient a, b et c des entiers, avec $a \neq 0$. Démontrer les énoncés suivants.

a) Si $a \mid b$, alors $a \mid bc$.
b) Si $b \neq 0$, alors $a \mid b$ et $b \mid c$ entraînent $a \mid c$.
c) Si $a \mid b$ et $a \mid c$, alors $a \mid (bx + cy)$ pour tous les entiers x et y.
d) Si $b \neq 0$, alors $a \mid b$ et $b \mid a$ entraînent $a = \pm b$.
e) Si $c \neq 0$ et $a \mid b$, alors $ac \mid bc$.

Exercice E.2.

a) Effectuer la division euclidienne de 250 par 13, puis celle de -250 par 13.
b) Montrer que le carré d'un entier est toujours congru à 0 ou à 1 modulo 4. On pourra distinguer les cas selon le reste de la division de l'entier par 2.
c) En déduire qu'aucun entier de la forme $4k + 3$ n'est une somme de deux carrés d'entiers.

Exercice E.3.

a) À l'aide du lemme E.3, déterminer si 221 et 223 sont premiers. Combien de divisions faut-il effectuer, au plus, dans chaque cas ?
b) Factoriser 6409 en nombres premiers, en justifiant pourquoi on peut s'arrêter.
c) Soit $n \geq 2$. Montrer que les $n-1$ entiers consécutifs $n!+2, n!+3, \ldots, n!+n$ sont tous composés. Qu'en déduit-on sur les écarts entre nombres premiers consécutifs ?

Exercice E.4. Le but de cet exercice est d'adapter la démonstration d'Euclide pour montrer qu'il existe une infinité de nombres premiers de la forme $4k + 3$.

a) Montrer qu'un nombre premier impair est congru à 1 ou à 3 modulo 4.
b) Montrer que le produit de deux entiers congrus à 1 modulo 4 est congru à 1 modulo 4. En déduire qu'un produit d'un nombre quelconque de tels entiers est aussi congru à 1 modulo 4.

c) Supposons qu'il n'y ait qu'un nombre fini de nombres premiers de la forme $4k+3$, soit $p_1, \dots, p_r$, et posons $N = 4p_1p_2\cdots p_r - 1$. Montrer que N possède un facteur premier de la forme $4k+3$, et conclure.

d) Pourquoi l'argument ne fonctionne-t-il pas pour les nombres premiers de la forme $4k+1$?

Exercice E.5. À l'aide de l'algorithme d'Euclide et de calculs supplémentaires, répondre aux questions suivantes.

a) Trouver $\text{pgcd}(42\,823, 6409)$.

b) Trouver des entiers x et y tels que $42\,823\,x + 6409\,y = 17$.

c) Factoriser 42 823 et 6409 en nombres premiers (pour 6409, voir l'exercice E.3).

d) Retrouver le PGCD à partir des factorisations.

Exercice E.6. Un artisan veut carreler un plancher rectangulaire de 3,15 m sur 1,35 m avec des tuiles carrées identiques, sans en découper aucune.

a) Quelle est la plus grande dimension possible des tuiles, si leur côté est un nombre entier de centimètres ? Combien en faudra-t-il ?

b) Dessiner, à la manière de la figure E.1, le découpage d'Euclide d'un rectangle 315×135, et faire le lien avec les divisions de l'algorithme d'Euclide.

c) L'artisan dispose en réalité de tuiles de 20 cm de côté. Peut-il carreler le plancher sans découpe ? Justifier à l'aide de la proposition E.8.

Exercice E.7.

a) Calculer $\text{pgcd}(360, 84)$ et $\text{ppcm}(360, 84)$ à l'aide des factorisations, et vérifier que leur produit vaut 360×84.

b) Trouver tous les couples d'entiers positifs (a, b), avec $a \leq b$, tels que $\text{pgcd}(a, b) = 12$ et $\text{ppcm}(a, b) = 360$.

Exercice E.8.

a) Montrer que, pour tout entier n, les entiers n et $n+1$ sont premiers entre eux, ainsi que $2n+1$ et $3n+1$.

b) L'équation $15x + 21y = 10$ a-t-elle des solutions entières ?

c) Trouver toutes les solutions entières de $15x + 21y = 9$.

d) On dispose d'un récipient de 5 L et d'un récipient de 7 L, sans graduation, ainsi que d'un robinet et d'un évier. Expliquer comment l'égalité $3 \times 5 - 2 \times 7 = 1$ suggère une façon d'obtenir exactement 1 L d'eau.

Exercice E.9. Soient a, b, c, d, x et y des entiers, et $m \geq 1$ un entier. Démontrer les énoncés suivants.

a) Les énoncés $a \equiv b \pmod m$, $b \equiv a \pmod m$ et $a - b \equiv 0 \pmod m$ sont équivalents.

b) Si $a \equiv b \pmod m$ et $c \equiv d \pmod m$, alors $a + c \equiv b + d \pmod m$.

c) Si $a \equiv b \pmod m$ et $b \equiv c \pmod m$, alors $a \equiv c \pmod m$.

d) Si $ax \equiv ay \pmod m$ et $\text{pgcd}(a, m) = 1$, alors $x \equiv y \pmod m$.

e) En général, $ax \equiv ay \pmod m$ n'entraîne pas $x \equiv y \pmod m$.

Exercice E.10. On note $\overline{c_m \cdots c_1 c_0}$ l'entier dont l'écriture décimale est formée des chiffres $c_m, \dots, c_1, c_0$.

a) Montrer que $\overline{c_m \cdots c_1 c_0} \equiv c_0 - c_1 + c_2 - \cdots + (-1)^m c_m \pmod{11}$. En déduire un critère de divisibilité par 11.
b) Les nombres 123 456 789 et 918 082 sont-ils divisibles par 9 ? par 11 ?
c) Montrer qu'un nombre de quatre chiffres de la forme $\overline{abba}$ (un palindrome, comme 3443) est toujours divisible par 11.

Exercice E.11. Un code ISBN-10, utilisé pour identifier les livres jusqu'en 2007, est une suite de dix chiffres $x_1 x_2 \cdots x_{10}$ (le dernier pouvant valoir 10, noté X) qui vérifie

$$1 \cdot x_1 + 2 \cdot x_2 + 3 \cdot x_3 + \cdots + 10 \cdot x_{10} \equiv 0 \pmod{11}.$$

a) Trouver le chiffre de contrôle x_{10} du code dont les neuf premiers chiffres sont 2 7 6 5 4 0 9 1 2. On pourra remarquer que $10 \equiv -1 \pmod{11}$.
b) Montrer que si l'on modifie un seul chiffre d'un code valide, le code obtenu n'est plus valide. Quel résultat de l'annexe utilise-t-on ?
c) Montrer que si l'on échange deux chiffres consécutifs différents d'un code valide, le code obtenu n'est plus valide.
d) Pourquoi a-t-on choisi le module 11 plutôt que 10, qui aurait évité le symbole X ?

Exercice E.12. Résoudre chacune des congruences suivantes, sans calculatrice, ou justifier qu'elle n'a pas de solution.

a) $15x \equiv 2 \pmod{4}$
b) $12x \equiv 4 \pmod{3}$
c) $2x \equiv 3 \pmod{6}$
d) $2x \equiv 3 \pmod{7}$
e) $6x \equiv 4 \pmod{10}$

Exercice E.13. Trouver, s'ils existent, les inverses suivants.

a) L'inverse de 12 modulo 23.
b) L'inverse de 33 modulo 114.
c) L'inverse de 101 modulo 102.
d) L'inverse de 17 modulo 43, à l'aide de l'algorithme d'Euclide étendu.

Exercice E.14.

a) Calculer le reste de la division de 3^{100} par 7.
b) Soit p un nombre premier. Montrer que

$$2^{p-1} + 3^{p-1} + \cdots + (p-1)^{p-1} \equiv -2 \pmod{p}.$$

c) Vérifier que $2^{10} \equiv 1 \pmod{341}$, et en déduire que $2^{340} \equiv 1 \pmod{341}$. Que peut-on en conclure sur la réciproque du petit théorème de Fermat, sachant que $341 = 11 \times 31$?

Exercice E.15.

a) Montrer que, pour tout entier n, $n^5 - n$ est divisible par 30.
b) Montrer que, pour tout entier n, $n^2(n^4 - 1)$ est divisible par 60.

Exercice E.16.

a) Sans utiliser la proposition E.23, montrer que, pour un nombre premier p et un entier $k \geq 1$, $\varphi(p^k) = p^k - p^{k-1}$.
b) Calculer $\varphi(12)$, $\varphi(100)$ et $\varphi(360)$.
c) Dresser la liste des entiers de 1 à 12 premiers avec 12, et vérifier le théorème d'Euler pour chacun d'eux en calculant son carré modulo 12.

Exercice E.17.

a) Énoncer le théorème d'Euler.
b) Trouver les deux derniers chiffres de 7^{222}.
c) Trouver les trois derniers chiffres de 43^{402}, à l'aide d'une calculatrice au besoin, mais pas d'un ordinateur.

Exercice E.18. Résoudre les systèmes de congruences suivants, sans calculatrice.

a) $x \equiv 1 \pmod 2$, $x \equiv 1 \pmod 3$, $x \equiv 1 \pmod 4$. Les modules sont-ils premiers entre eux deux à deux ? Comment procéder ?
b) $2x \equiv 3 \pmod 5$ et $3x \equiv 1 \pmod 8$.
c) Une paysanne apporte des œufs au marché. Si elle les compte par deux, par trois, par quatre, par cinq ou par six, il en reste toujours un ; si elle les compte par sept, il n'en reste aucun. Quel est le plus petit nombre d'œufs possible ?

Exercice E.19. Plusieurs notions de cette annexe sont enseignées, sous une forme plus élémentaire, au primaire et au secondaire : la division avec reste, les critères de divisibilité, les nombres premiers, le PGCD et le PPCM.

a) Un élève du premier cycle calcule le PGCD de 84 et de 126 en factorisant les deux nombres. Montrer comment l'algorithme d'Euclide, présenté avec la figure E.1, permettrait d'obtenir le même résultat. Quels avantages et quelles limites voyez-vous à chaque méthode en classe ?
b) La « preuve par neuf » permet de contrôler une multiplication : on remplace chaque facteur et le résultat par la somme de leurs chiffres, réduite modulo 9, et on vérifie que l'égalité tient encore. Expliquer pourquoi la méthode fonctionne, puis montrer par un exemple qu'elle ne détecte pas toutes les erreurs.
c) Proposer une activité, accessible à des élèves du secondaire, qui fasse découvrir la démonstration d'Euclide sur l'infinitude des nombres premiers.

Solutions et réponses

Chapitre 1 – Les transformations du plan

Exercice 1.1. Les cas (a), (d), (e), (f) et (i) satisfont la définition. En (b), f_2 n'est pas définie lorsque $x = \pi/2 + k\pi$: ce n'est pas une correspondance partout définie. En (c), f_3 n'est ni injective ($f_3(x, y) = f_3(-x, y)$) ni surjective (la première coordonnée de l'image est toujours ≥ 0). Le cas (g) présente les mêmes défauts que (c), pour les mêmes raisons. En (h), f_8 est injective, mais pas surjective : $\arctan y \in]-\pi/2, \pi/2[$ pour tout y, si bien qu'aucun point de seconde coordonnée $\geq \pi/2$ n'est atteint. Dans ce dernier cas, seule la surjectivité fait défaut. Parmi les transformations obtenues, (d), (e), (f) et (i) sont affines, alors que (a) ne l'est pas.

Exercice 1.2.

a) **Faux** : deux points distincts ont des images distinctes par injectivité.

b) **Faux** : chaque point possède un antécédent par surjectivité.

c) **Vrai** : l'égalité $f \circ g = \mathrm{id}$ avec g bijective implique que f est la réciproque de g. La composition d'une bijection et de sa réciproque donne l'identité dans les deux ordres.

d) **Faux**. Contre-exemple : $f_1(x, y) = (x^5, y)$ de l'exercice 1.1 est une transformation du plan, mais elle envoie les points alignés $(0, 0)$, $(1, 1)$, $(2, 2)$ sur $(0, 0)$, $(1, 1)$, $(32, 2)$, qui ne sont plus alignés.

e) **Vrai**, par la proposition 1.3.

f) **Faux** : c'est vrai pour une transformation *affine* (théorème 1.15, avec $Q_i = P_i$). Cette hypothèse n'est toutefois pas donnée ici. Pour construire un contre-exemple, on peut fixer les trois points non alignés et échanger deux autres points du plan, en laissant les autres inchangés. On obtient encore une bijection.

Exercice 1.3. Soient $f, g \in \mathcal{T}$. La composée $f \circ g$ est partout définie et univoque comme composée de fonctions. Pour l'injectivité, supposons que $f(g(P)) = f(g(Q))$, l'injectivité de f donne $g(P) = g(Q)$, puis celle de g donne $P = Q$. Pour la surjectivité, considérons $R \in \Pi$; par surjectivité de f, il existe S tel que $f(S) = R$, puis, par surjectivité de g, il existe P tel que $g(P) = S$; alors $(f \circ g)(P) = R$.

Exercice 1.4.

a) Le parallélogramme plein de sommets $(-1, 1)$, $(1, 5)$, $(0, 5)$ et $(-2, 1)$.

b) Le triangle plein de sommets $(-1, 3)$, $(-1, 1)$ et $(-3, -1)$.

c) $\det(A) = (1)(0) - (-1)(2) = 2$: l'aire de chacune des images est donc le double de celle de la figure de départ. Le signe positif du déterminant indique aussi que l'orientation est conservée.

Exercice 1.5. ($\Rightarrow$) Supposons w similitude de rapport k et soit U unitaire. Comme w est linéaire, $w(\mathbf{0}) = \mathbf{0}$, donc

$$\|w(U)\| = \|w(U) - w(\mathbf{0})\| = k\,\|U - \mathbf{0}\| = k\,\|U\| = k.$$

($\Leftarrow$) Supposons $\|w(U)\| = k$ pour tout U unitaire, et soient $P, Q \in \mathbb{R}^2$. Si $P = Q$, l'égalité (1.4) est immédiate. Sinon, posons

$$V = \frac{P - Q}{\|P - Q\|},$$

qui est unitaire, de sorte que $\|w(V)\| = k$. La linéarité permet alors d'écrire,

$$w(V) = \frac{1}{\|P - Q\|}\, w(P - Q) = \frac{1}{\|P - Q\|}\, [w(P) - w(Q)],$$

d'où

$$k = \|w(V)\| = \frac{1}{\|P - Q\|}\, \|w(P) - w(Q)\|,$$

c'est-à-dire $\|w(P) - w(Q)\| = k\,\|P - Q\|$.

Exercice 1.6. La proposition 1.8 permet de se limiter, dans chaque cas, à la partie linéaire. Translation : $A = I$, et $\|P - Q\| = \|P - Q\|$, donc $k = 1$. Homothétie de rapport c : $\|cP - cQ\| = |c|\,\|P - Q\|$, donc $k = |c|$. Rotation et réflexion : leurs matrices vérifient $A^\top A = I$, donc $\|AP - AQ\|^2 = (P - Q)^\top A^\top A (P - Q) = \|P - Q\|^2$ et $k = 1$.

Exercice 1.7. C'est le point $\big(cx + h(1 - c),\ cy + k(1 - c)\big)$. Voir la proposition 1.5. On peut vérifier que le centre (h, k) est fixe, puisque $ch + h - ch = h$.

Exercice 1.8. La partie linéaire est $r_{-3\pi/4}$, et la partie translation se déduit de l'image de l'origine :

$$\begin{aligned} w\begin{pmatrix} x \\ y \end{pmatrix} &= \begin{pmatrix} \cos(-\frac{3\pi}{4}) & -\sin(-\frac{3\pi}{4}) \\ \sin(-\frac{3\pi}{4}) & \cos(-\frac{3\pi}{4}) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 1 - \frac{\sqrt{2}}{4} \\ \frac{\sqrt{2}}{4} \end{pmatrix} \\ &= \begin{pmatrix} -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 1 - \frac{\sqrt{2}}{4} \\ \frac{\sqrt{2}}{4} \end{pmatrix}. \end{aligned}$$

Exercice 1.9. $w\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -4 & 3 \\ 2 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 37 \\ -43 \end{pmatrix}$. Voir l'exemple 1.10 pour la démarche, et l'annexe A pour le code.

Exercice 1.10. Le segment de départ mesure $2\sqrt{2}$, celui d'arrivée $\sqrt{2}$: le rapport est nécessairement $k = \frac{1}{2}$. Une première solution s'obtient en composant la translation qui envoie P sur l'origine, une rotation de $\pi/2$, l'homothétie de rapport $\frac{1}{2}$, puis la translation qui ramène l'origine sur A :

$$w_1\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 & -1/2 \\ 1/2 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 5/2 \\ 5/2 \end{pmatrix}.$$

Une seconde s'obtient en remplaçant la rotation par une réflexion :

$$w_2\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -1/2 & 0 \\ 0 & 1/2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 3/2 \\ 3/2 \end{pmatrix}.$$

Ces deux solutions sont les seules possibles : la partie linéaire d'une similitude de rapport k est nécessairement de la forme $k\,r_\theta$ ou $k\,R_\theta$ (section 1.8), et les conditions $w(P) = A$, $w(Q) = B$ déterminent alors complètement k, θ et la translation dans chaque cas. Comparez avec la remarque 1.11.

Exercice 1.11. $w\begin{pmatrix}x\\y\end{pmatrix} = \frac{1}{2}\begin{pmatrix}1 & 1\\1 & 1\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}$. Son déterminant est nul, donc ce n'est pas une transformation du plan : l'image du plan est contenue dans la droite $y = x$. Voir l'exemple 1.7.

Exercice 1.12. Si $(u, v) = w(x, y)$, alors $x = (u - e)/a$ et $y = (v - f)/b$, d'où

$$\frac{(u-e)^2}{(ar)^2} + \frac{(v-f)^2}{(br)^2} = 1 :$$

Il s'agit d'une ellipse de centre (e, f), d'axe horizontal de longueur $2ar$ et d'axe vertical de longueur $2br$. En particulier, l'image est un cercle si et seulement si $a = b$, c'est-à-dire si et seulement si w est une similitude.

Exercice 1.13. Une telle similitude n'existe pas. En effet, une similitude multiplie toutes les distances par un même facteur k ; elle conserve donc le rapport des longueurs des demi-axes. Ce rapport vaut 1 pour le cercle et 2 pour l'ellipse.

Exercice 1.14.

a) Prendre $A = \begin{pmatrix}1/2 & 0\\0 & 3/2\end{pmatrix}$, de déterminant $3/4 < 1$. Pour vérifier qu'il ne s'agit pas d'une contraction, considérons $P = (0, 0)$ et $Q = (0, 1)$ ont pour images $(0, 0)$ et $(0, 3/2)$, dont la distance a augmenté.

b) Par la proposition 1.10, $|\det(A)| = k^2$. Donc $|\det(A)| < 1 \iff k^2 < 1 \iff 0 < k < 1$, ce qui est la définition d'une contraction.

Exercice 1.15.

a) Les trois triangles ont un angle droit et partagent chacun un angle aigu avec $\mathcal{T}$; ils sont donc semblables par le critère AA.

b) Le côté $\overline{AC} = b$ de $\mathcal{T}$ correspond à l'hypoténuse a de $\mathcal{T}$ dans la similitude ; on trouve $k_1 = b/a$ et $k_2 = c/a$.

c) Par la proposition 1.11, $s_1 = k_1^2 s$ et $s_2 = k_2^2 s$. Alors

$$s = s_1 + s_2 = (k_1^2 + k_2^2)\, s \implies 1 = \frac{b^2}{a^2} + \frac{c^2}{a^2} \implies a^2 = b^2 + c^2.$$

d) $\mathcal{E} = \mathcal{T}$ convient, puisque $\mathcal{T} = \mathcal{T}_1 \cup \mathcal{T}_2$. On obtient ainsi un exemple d'*ensemble fixe* d'un système de fonctions itérées : un triangle rectangle est réunion de deux copies réduites de lui-même.

Exercice 1.16. $(t \circ R)\begin{pmatrix}x\\y\end{pmatrix} = \begin{pmatrix}1 & 0\\0 & -1\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix} + \begin{pmatrix}3\\0\end{pmatrix}$. La partie linéaire est R_0 (déterminant -1), et la translation $(3, 0)$ est déjà parallèle à l'axe Ox : c'est une réflexion glissée d'axe Ox et de vecteur de glissement $(3, 0)$, puisque le vecteur de translation n'a pas de composante perpendiculaire à l'axe. Avec le vecteur $(0, 3)$, la translation est au contraire entièrement perpendiculaire à l'axe : son effet se traduit par un déplacement de l'axe, et $t \circ R$ est une réflexion pure, d'axe $y = 3/2$.

Exercice 1.17. Nous utilisons la propriété suivante : une matrice orthogonale 2×2 est nécessairement la matrice d'une rotation ou d'une réflexion (section 1.8). Après avoir vérifié que $A^\top A = k^2 I$ avec $k = \sqrt{(A^\top A)_{11}}$, on pose $Q = A/k$, qui est orthogonale. Si $\det(Q) > 0$, la matrice Q est une rotation et $\theta = \operatorname{atan2}(Q_{21}, Q_{11})$. Sinon, Q est une réflexion et l'angle de son axe est $\frac{1}{2}\operatorname{atan2}(Q_{21}, Q_{11})$, d'après l'équation (1.2). Le code complet est donné à l'annexe A.

Exercice 1.18. (i) Ce qui est juste : *toutes les longueurs* sont bien multipliées par 3 : les côtés, les diagonales, le périmètre. L'erreur est d'étendre cette règle à l'aire. (ii) Contre-exemple : un rectangle 2×5 a une aire de 10 ; son image est un rectangle 6×15, d'aire 90, soit 9 fois plus et non 3. (iii) Explication : l'aire est un produit de deux longueurs. Chacune est multipliée par 3, donc le produit est multiplié par $3 \times 3 = 9$. Pour l'illustrer, on peut effectuer un découpage : le rectangle image se pave exactement par 9 copies du rectangle de départ. Le même argument donne $3^3 = 27$ pour les volumes (voir la remarque 1.6).

Exercice 1.19. Plusieurs réponses sont possibles ; en voici une.

a) *Le facteur k^2 pour les aires* (proposition 1.11) : en troisième secondaire, il se présente par un pavage du rectangle image par k^2 copies du rectangle de départ (remarque 1.6, exercice 1.18). *La composée de deux réflexions* (théorème 1.13) : en première secondaire, l'activité des deux miroirs fait constater que deux réflexions d'axes sécants donnent une rotation d'angle double, et que deux réflexions d'axes parallèles donnent une translation. L'enseignant qui sait que toute isométrie est une composée de réflexions (remarque 1.8) peut donner aux élèves une vision unifiée des isométries, plutôt qu'une liste de quatre transformations sans lien entre elles.

b) Pour le facteur k^2, l'erreur prévisible est l'extension abusive de la linéarité : « tout est multiplié par k ». Le chapitre montre que les longueurs, les aires et les volumes ne sont pas des grandeurs de même nature ; l'enseignant peut donc proposer d'emblée un contre-exemple numérique plutôt que d'attendre que l'erreur se manifeste en évaluation.

c) Oui, à condition que l'outil serve à vérifier une prédiction formulée à l'avance (remarque 1.11). Écrire la fonction qui calcule l'image d'un point oblige à expliciter la règle de la transformation et à la concevoir comme une correspondance qui agit sur *tous* les points du plan, et non seulement sur les sommets de la figure dessinée. Le temps d'apprentissage de l'outil doit toutefois rester modeste par rapport au temps consacré aux mathématiques.

Chapitre 2 – Espaces métriques

Exercice 2.1. La positivité, la séparation et la symétrie se vérifient à partir des propriétés de la fonction : $t \mapsto t/(1+t)$ est nulle seulement en 0 et positive ailleurs. Pour l'inégalité triangulaire, posons $a = |x - z|$, $b = |z - y|$ et $c = |x - y|$; on sait que $c \leq a + b$. Puisque $t \mapsto t/(1+t)$ est croissante,

$$\begin{aligned}\delta(x,y) = \frac{c}{1+c} \leq \frac{a+b}{1+a+b} &= \frac{a}{1+a+b} + \frac{b}{1+a+b} \\ &\leq \frac{a}{1+a} + \frac{b}{1+b} = \delta(x,z) + \delta(z,y).\end{aligned}$$

Contrairement à d, la métrique δ est bornée : $\delta(x,y) < 1$ pour tout x, y.

Exercice 2.2.

a) $d_{\text{Ham}}(0000, 0110) = 2, d_{\text{Ham}}(0000, 1011) = 3, d_{\text{Ham}}(0000, 1101) = 3, d_{\text{Ham}}(0110, 1011) = 3$, $d_{\text{Ham}}(0110, 1101) = 3$, $d_{\text{Ham}}(1011, 1101) = 2$. La distance minimale entre deux messages valides distincts est donc 2.

b) Non : la note historique 2.1 exige une distance minimale d'au moins 3 pour corriger automatiquement toute erreur d'un bit, alors qu'ici la distance minimale entre messages valides n'est que 2. On peut le vérifier sur l'exemple suivant : si une erreur inverse le deuxième bit de 0000, le message reçu est 0100. Or $d_{\text{Ham}}(0100, 0000) = 1$ *et* $d_{\text{Ham}}(0100, 0110) = 1$: le message reçu est aussi près de 0000 que de 0110, de sorte qu'on ne peut pas déterminer lequel a été envoyé. Une distance minimale de 2 suffit à *détecter* qu'une erreur s'est produite (un message corrompu par une seule erreur ne peut jamais devenir un autre message valide), mais pas à la corriger automatiquement.

Exercice 2.3.

a) A_1 est ouvert (c'est une boule ouverte) et borné, mais non fermé : son complément n'est pas ouvert, car tout point du cercle unité, lui-même dans le complément, n'admet aucune boule qui évite A_1. N'étant pas fermé, A_1 n'est pas compact.

b) A_2 est fermé et borné, donc compact (théorème 2.1) ; il n'est pas ouvert (le point $(1, 0)$, sur le bord, n'admet aucune boule ouverte contenue dans A_2).

c) A_3 n'est ni ouvert ni fermé : l'origine est un point limite de A_3 non contenu dans A_3 (le théorème 2.1 ne s'applique donc pas, puisque A_3 n'est pas fermé) ; il n'est donc pas compact non plus.

d) $A_4 = \mathbb{R}^2$ est à la fois ouvert et fermé (remarque 2.3), mais non borné, donc non compact.

Exercice 2.4. Puisque $A = \{0\}$, $\sup_{a\in A} d(a, B) = d(0, B) = 0$, car $0 \in B$. Pour calculer le second terme, $\sup_{b\in B} d(b, A) = \sup_{b\in[-1,1]} |b| = 1$, atteint en $b = \pm 1$. Ainsi $d_H(A, B) = \max\{0, 1\} = 1$.

Exercice 2.5.

a) $|f(x) - f(y)| = \frac{1}{3}|x - y|$, donc f est une contraction de rapport $k = 1/3$.

b) $\overline{x} = \frac{1}{3}\overline{x} + 2 \implies \frac{2}{3}\overline{x} = 2 \implies \overline{x} = 3$.

c) $x_1 = f(0) = 2$, donc $d(x_0, x_1) = 2$. La proposition 2.7 donne $d(x_3, \overline{x}) \leq \dfrac{(1/3)^3}{1 - 1/3} \cdot 2 = \dfrac{1/27}{2/3} \cdot 2 = \dfrac{2}{18} = \frac{1}{9} \approx 0{,}111$.

Exercice 2.6.

a) $g(x) = x - \dfrac{f(x)}{f'(x)} = x - \dfrac{x^2 - 3}{2x} = \dfrac{2x^2 - (x^2 - 3)}{2x} = \dfrac{x^2 + 3}{2x} = \dfrac{1}{2}\left(x + \dfrac{3}{x}\right)$.

b) $x_1 = g(2) = \frac{1}{2}(2 + 1{,}5) = 1{,}75$, une seule décimale exacte (1,7). $x_2 = g(1{,}75) = \frac{1}{2}(1{,}75 + 3/1{,}75) \approx 1{,}732\,142\,857$, déjà exact à 4 décimales près ($\sqrt{3} \approx 1{,}732\,050\,8$) : le nombre de décimales exactes a approximativement doublé d'une itération à l'autre, ce qui illustre la convergence quadratique présentée à la section 2.7.1.

Exercice 2.7. Plusieurs réponses sont possibles ; en voici une.

a) On place les élèves dans un quartier en damier : un taxi doit aller d'un coin de rue à un autre. On leur demande de calculer la longueur du trajet le plus court, puis de

tracer tous les points situés à 3 coins de rue d'un point donné. Ils obtiennent un carré posé sur sa pointe, et non un cercle. L'objectif n'est pas d'enseigner une nouvelle formule, mais de faire comprendre que la formule $\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$ découle du théorème de Pythagore et décrit un déplacement « à vol d'oiseau » : elle correspond à un choix de modèle. L'activité permet aussi de réinvestir la valeur absolue.

b) C'est la méthode de Héron (ou méthode babylonienne). Pour approcher $\sqrt{10}$, on part de $x_0 = 3$; si x_n est trop petit, $10/x_n$ est trop grand, et leur moyenne est une meilleure approximation. Les élèves calculent quelques termes à la calculatrice ou dans un tableur et constatent que les décimales se stabilisent rapidement. L'enseignant mobilise la notion de point fixe : $\sqrt{a}$ est la valeur qui reste inchangée par l'itération, et la suite converge vers elle (section 2.7.1). L'activité donne aussi un sens au caractère irrationnel de $\sqrt{10}$, dont le développement décimal ne se termine pas.

c) Pour l'élève, une distance est une mesure physique, prise avec une règle. S'il apprend qu'il existe « plusieurs distances », il peut croire que la formule apprise est fausse ou arbitraire. Il faut donc préciser que chaque distance répond à une question différente (la longueur du trajet d'un piéton, d'un oiseau ou du nombre de bits erronés), et que la distance euclidienne reste celle de la géométrie étudiée en classe.

Chapitre 3 – Fractales et systèmes itératifs de fonctions

Exercice 3.1.

a) À chaque étape, chaque segment du contour est remplacé par quatre segments trois fois plus courts : le périmètre est multiplié par $\frac{4}{3}$. Le périmètre à l'étape n vaut donc $P_n = 3\left(\frac{4}{3}\right)^n$, qui tend vers l'infini ; le périmètre du flocon est infini.

b) Notons $A_0 = \frac{\sqrt{3}}{4}$ l'aire du triangle de départ. À l'étape $n \geq 1$, on ajoute un petit triangle équilatéral sur chacun des $3 \cdot 4^{n-1}$ segments de l'étape précédente ; ces triangles ont un côté de $\left(\frac{1}{3}\right)^n$, donc une aire de $\left(\frac{1}{9}\right)^n A_0$. En additionnant les aires ajoutées, nous obtenons

$$\begin{aligned} A &= A_0 + \sum_{n=1}^{\infty} 3 \cdot 4^{n-1} \left(\frac{1}{9}\right)^n A_0 = A_0 \left(1 + \frac{3}{4} \sum_{n=1}^{\infty} \left(\frac{4}{9}\right)^n\right) \\ &= A_0 \left(1 + \frac{3}{4} \cdot \frac{4/9}{1 - 4/9}\right) = \frac{8}{5} A_0. \end{aligned}$$

Pour un triangle de côté 1, l'aire du flocon est donc $\frac{8}{5} \cdot \frac{\sqrt{3}}{4} = \frac{2\sqrt{3}}{5} \approx 0{,}693$.

Exercice 3.2.

a) À chaque étape, chaque carré est remplacé par 5 carrés dont l'aire est 9 fois plus petite : l'aire est multipliée par $\frac{5}{9}$. Puisque $A_1 = a^2$, on a $A_n = \left(\frac{5}{9}\right)^{n-1} a^2$, qui tend vers 0.

b) Le flocon est formé de $N = 5$ copies de lui-même, de rapport $r = \frac{1}{3}$, dont les seuls points communs sont des sommets. Sa dimension de similitude est $\ln 5 / \ln 3 \approx 1{,}465$.

Exercice 3.3.

a) À la première étape, on découpe $[0,1]$ en dix intervalles de longueur $\frac{1}{10}$, et on retire ceux où la première décimale est impaire : il reste les cinq intervalles $[0\,;0{,}1]$, $[0{,}2\,;0{,}3]$, …, $[0{,}8\,;0{,}9]$ (on sépare les bornes par un point-virgule pour éviter la confusion avec la virgule décimale). À l'étape n, on découpe de même chacun des intervalles restants en dix, et on retire ceux où la n^{e} décimale est impaire. Il reste 5^n intervalles de longueur 10^{-n}, et E est l'intersection de toutes les étapes, comme dans la construction de l'ensemble de Cantor (remarque 3.1).

b) Les nombres de E dont la première décimale est $j \in \{0,2,4,6,8\}$ s'obtiennent à partir d'un nombre de E en déplaçant ses décimales d'un rang vers la droite, puis en insérant j à la première position : ils forment l'ensemble $w_j(E)$, où $w_j(x) = \frac{1}{10}x + \frac{j}{10}$. Ainsi $E = w_0(E) \cup w_2(E) \cup w_4(E) \cup w_6(E) \cup w_8(E)$ est autosimilaire, formé de 5 copies de rapport $\frac{1}{10}$. Sa dimension de similitude est $\ln 5 / \ln 10 \approx 0{,}699$.

Exercice 3.4.

a) L'aire de l'étape n est $\left(\frac{3}{4}\right)^n \frac{\sqrt{3}}{4}$, qui tend vers 0 : l'aire du triangle de Sierpiński est nulle. Sa dimension de similitude est $\ln 3 / \ln 2 \approx 1{,}585$ (3 copies de rapport $\frac{1}{2}$).

b) L'aire de l'étape n est $\left(\frac{8}{9}\right)^n$, qui tend vers 0 : l'aire du tapis est nulle. Sa dimension de similitude est $\ln 8 / \ln 3 \approx 1{,}893$ (8 copies de rapport $\frac{1}{3}$).

Exercice 3.5.

a) À chaque étape, on conserve $27 - 7 = 20$ cubes sur 27 : le volume de l'étape n est $\left(\frac{20}{27}\right)^n$, qui tend vers 0. Le volume de l'éponge est nul, et sa dimension de similitude est $\ln 20 / \ln 3 \approx 2{,}727$.

b) Chaque tétraèdre conservé a des arêtes deux fois plus courtes, donc un volume $2^3 = 8$ fois plus petit (remarque 1.6) ; on en conserve 4, et le volume est multiplié par $\frac{4}{8} = \frac{1}{2}$ à chaque étape. Le volume est nul, et la dimension de similitude est $\ln 4 / \ln 2 = 2$. Cet exemple donne donc une fractale de dimension entière : elle a la dimension d'une surface sans en être une.

Exercice 3.6. Appliquons la formule (2.1).

a) $d_H(A,B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$: pour deux singletons, la distance de Hausdorff est la distance euclidienne.

b) Le point de B le plus éloigné de $(1,1)$ est le point du cercle diamétralement opposé, $(-\frac{r}{\sqrt{2}}, -\frac{r}{\sqrt{2}})$, à distance $r + \sqrt{2}$; la distance de $(1,1)$ à B est, elle, $\max\{0, \sqrt{2} - r\}$, qui est plus petite. Ainsi $d_H(A,B) = r + \sqrt{2}$.

c) Les points de B les plus éloignés de A sont $(0, \pm 2)$, à distance 2 de A ; les points de A les plus éloignés de B sont $(\pm 1, 0)$, à distance 1 de B. Ainsi $d_H(A,B) = 2$.

d) Posons $\delta = \sqrt{(h_2 - h_1)^2 + (k_2 - k_1)^2}$, la distance entre les centres. Le point de B le plus éloigné de A est sur la droite des centres, du côté opposé à A, à distance $\max\{0, \delta + r_2 - r_1\}$ de A ; symétriquement, les points de A sont à distance au plus $\max\{0, \delta + r_1 - r_2\}$ de B. Le plus grand de ces deux nombres est $d_H(A,B) = \delta + |r_2 - r_1|$.

Exercice 3.7.

a) Le rapport de contraction est le plus grand des rapports, soit $\frac{1}{2}$.

b) L'équation de Moran s'écrit $2\left(\frac{1}{2}\right)^d + 3\left(\frac{1}{4}\right)^d = 1$. En posant $u = \left(\frac{1}{2}\right)^d$, elle devient $3u^2 + 2u - 1 = 0$, c'est-à-dire $(3u-1)(u+1) = 0$. Comme $u > 0$, $u = \frac{1}{3}$, d'où $2^d = 3$ et $d = \ln 3 / \ln 2 \approx 1{,}585$.

c) L'équation $2\left(\frac{1}{2}\right)^d + 3\left(\frac{1}{3}\right)^d = 1$ ne se résout pas par les méthodes algébriques usuelles. La méthode de Newton (ou une simple recherche par dichotomie) en donne une valeur approchée : $d \approx 1{,}788$.

Exercice 3.8. Dans chaque cas, on écrit $w(\mathbf{x}) = A\mathbf{x} + \mathbf{t}$, puis on utilise le quadrillage pour déterminer les images des points repères. Pour la partie b), on adapte l'un des deux programmes de la section 3.6.

(a) Le flocon de l'exercice 3.2 : cinq homothéties de rapport $\frac{1}{3}$, $w_i(\mathbf{x}) = \frac{1}{3}\mathbf{x} + \mathbf{t}_i$ avec $\mathbf{t}_i \in \{(0,0), (\frac{2}{3},0), (\frac{1}{3},\frac{1}{3}), (0,\frac{2}{3}), (\frac{2}{3},\frac{2}{3})\}$. La dimension de similitude et sa version généralisée valent toutes deux $\ln 5 / \ln 3 \approx 1{,}465$.

(b) Trois similitudes de rapport $\frac{1}{2}$: $w_1(\mathbf{x}) = \frac{1}{2}\mathbf{x}$, $w_2(\mathbf{x}) = \frac{1}{2}\mathbf{x} + (\frac{1}{2},0)$ et $w_3(\mathbf{x}) = \frac{1}{2}\, r_{\pi/2}\, \mathbf{x} + (\frac{1}{2},\frac{1}{2})$, la copie du coin supérieur gauche étant tournée d'un quart de tour dans le sens antihoraire. Les deux dimensions valent $\ln 3 / \ln 2 \approx 1{,}585$: la rotation de la copie modifie la forme de l'attracteur sans modifier sa dimension.

(c) $w_1(\mathbf{x}) = \frac{1}{2}\mathbf{x}$, $w_2(\mathbf{x}) = \frac{1}{2}\mathbf{x} + (\frac{1}{2},\frac{1}{2})$, $w_3(\mathbf{x}) = \frac{1}{4}\mathbf{x} + (\frac{3}{4},0)$ et $w_4(\mathbf{x}) = \frac{1}{4}\mathbf{x} + (0,\frac{3}{4})$. Puisque les rapports sont différents, on ne peut pas utiliser la définition initiale de la dimension de similitude. La dimension généralisée vérifie $2\left(\frac{1}{2}\right)^d + 2\left(\frac{1}{4}\right)^d = 1$, soit $2u^2 + 2u - 1 = 0$ avec $u = \left(\frac{1}{2}\right)^d$, d'où $u = \frac{\sqrt{3}-1}{2}$ et $d = \ln(1+\sqrt{3}) / \ln 2 \approx 1{,}450$.

(d) Quatre homothéties de rapport $\frac{1}{4}$, dont les images se trouvent dans les quatre coins : $w_i(\mathbf{x}) = \frac{1}{4}\mathbf{x} + \mathbf{t}_i$, avec $\mathbf{t}_i \in \{(0,0), (\frac{3}{4},0), (0,\frac{3}{4}), (\frac{3}{4},\frac{3}{4})\}$. Les deux dimensions valent $\ln 4 / \ln 4 = 1$: cette « poussière » de points isolés a la même dimension qu'un segment.

(e) $w_i(\mathbf{x}) = \begin{pmatrix} 1/3 & 0 \\ 0 & 1/2 \end{pmatrix} \mathbf{x} + \mathbf{t}_i$, avec $\mathbf{t}_i \in \{(0,0), (\frac{2}{3},0), (\frac{1}{3},\frac{1}{2})\}$. Les facteurs de réduction horizontale et verticale sont différents ; ces applications ne sont donc pas des similitudes : ni la dimension de similitude, ni la dimension généralisée ne sont définies (remarque 3.8).

(f) $w_1(\mathbf{x}) = \frac{1}{2}\mathbf{x}$, et trois homothéties de rapport $\frac{1}{3}$, $w_i(\mathbf{x}) = \frac{1}{3}\mathbf{x} + \mathbf{t}_i$, les vecteurs $\mathbf{t}_i$ étant $(\frac{2}{3},0)$, $(0,\frac{2}{3})$ et $(\frac{2}{3},\frac{2}{3})$. La dimension de similitude n'existe pas ; la dimension généralisée est la solution de $\left(\frac{1}{2}\right)^d + 3\left(\frac{1}{3}\right)^d = 1$, que l'on approche par la méthode de Newton. En partant de $d_0 = 1{,}5$, on obtient en trois itérations : $d \approx 1{,}4243$.

Exercice 3.9.

a) L'attracteur est contenu dans le rectangle $R = [0,16] \times [0,10]$, et il est formé de trois copies de lui-même, contenues dans les trois quarts de R autres que le quart supérieur gauche. Les deux copies du bas sont des réductions de moitié :

$$f_1\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}, \qquad f_2\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 8 \\ 0 \end{pmatrix}.$$

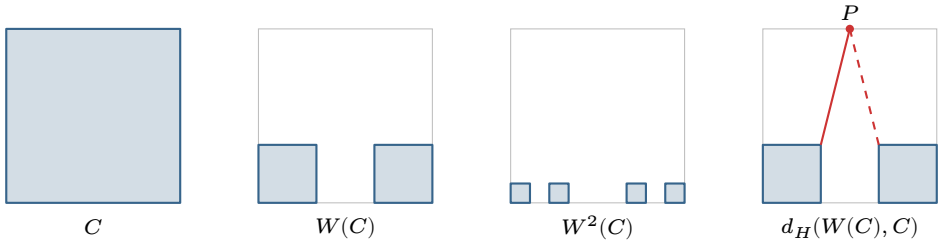
La copie du haut, contenue dans $[8,16]\times[5,10]$, est orientée horizontalement. Pour l'obtenir, on applique d'abord la rotation $r_{\pi/2} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, qui envoie R sur le rectangle $[-10,0] \times [0,16]$, puis la contraction non uniforme $c = \begin{pmatrix} 4/5 & 0 \\ 0 & 5/16 \end{pmatrix}$, qui ramène ce rectangle, de dimensions 10×16, aux dimensions 8×5, et enfin la translation t de vecteur $(16,5)$. Ainsi

$$f_3\begin{pmatrix} x \\ y \end{pmatrix} = c\, r_{\pi/2}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 16 \\ 5 \end{pmatrix} = \begin{pmatrix} 0 & -\frac{4}{5} \\ \frac{5}{16} & 0 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 16 \\ 5 \end{pmatrix}.$$

On vérifie par exemple que $f_3(0,0) = (16,5)$ et $f_3(16,10) = (8,10)$.

b) La dimension de similitude généralisée n'est pas définie, puisque f_3 n'est pas une similitude : elle réduit les distances horizontales d'un facteur $\frac{5}{16}$ et les distances verticales d'un facteur $\frac{4}{5}$. Les méthodes de ce chapitre ne permettent donc pas de calculer la dimension de A.

Exercice 3.10. a), b) et c) On a $W(C) = [0,\frac{1}{3}]\times[0,\frac{1}{3}] \cup [\frac{2}{3},1]\times[0,\frac{1}{3}]$, et $W^2(C)$ est formé de quatre carrés de côté $\frac{1}{9}$, posés sur l'axe des x et dont les bases sont $[0,\frac{1}{9}]$, $[\frac{2}{9},\frac{1}{3}]$, $[\frac{2}{3},\frac{7}{9}]$ et $[\frac{8}{9},1]$.



d) Puisque $W(C) \subseteq C$, il suffit de mesurer l'éloignement des points de C par rapport à $W(C)$. Le point de C le plus éloigné de $W(C)$ est $P = (\frac{1}{2},1)$, au milieu du côté supérieur (dernier panneau) ; les points de $W(C)$ les plus proches de lui sont $(\frac{1}{3},\frac{1}{3})$ et $(\frac{2}{3},\frac{1}{3})$. Ainsi $d_H(W(C),C) = \sqrt{(\frac{1}{6})^2 + (\frac{2}{3})^2} = \frac{\sqrt{17}}{6} \approx 0{,}687$. (Le coin $(0,1)$, par exemple, n'est qu'à distance $\frac{2}{3} = \frac{\sqrt{16}}{6}$ de $W(C)$.)

e) Les deux applications sont des homothéties de rapport $\frac{1}{3}$: le rapport de contraction est $k = \frac{1}{3}$.

f) Le théorème 3.3 donne $d_H(W^n(C),A) \leq \dfrac{(1/3)^n}{1-1/3}\cdot\dfrac{\sqrt{17}}{6} = \dfrac{\sqrt{17}}{4}\left(\dfrac{1}{3}\right)^n$. Ce majorant vaut environ $0{,}0127$ pour $n=4$ et $0{,}0042$ pour $n=5$: il suffit de choisir $n \geq 5$.

Exercice 3.11.

a) Avancer à mi-chemin vers le sommet S_i, c'est appliquer l'homothétie $h_{S_i,1/2}$. Le SIF est formé des quatre homothéties de rapport $\frac{1}{2}$ centrées aux sommets, qui envoient le carré unité C sur ses quatre quarts. On a donc $W(C) = C$, et par unicité, l'attracteur est le carré plein C : les points remplissent uniformément le carré, et l'attracteur obtenu n'est pas fractal.

b) Avancer des deux tiers du chemin vers S_i mène au point $P + \frac{2}{3}(S_i - P) = \frac{1}{3}P + \frac{2}{3}S_i$: c'est l'homothétie $h_{S_i,1/3}$. Les quatre copies de rapport $\frac{1}{3}$ sont les carrés de coin de côté $\frac{1}{3}$; l'attracteur est une « poussière de Cantor » formée de points isolés, de dimension de similitude $\ln 4/\ln 3 \approx 1{,}262$, la même que celle de la courbe de von Koch.

Exercice 3.12. Un SIF possible est celui du tableau de l'exemple 3.8. On le retrouve par la méthode du collage : on repère trois points de la fougère (la base de la tige, la pointe de la fronde et l'extrémité de la foliole inférieure gauche, par exemple) et leurs positions sur chacune des quatre pièces de la figure 3.11, puis on résout, pour chaque pièce, les systèmes du théorème 1.15. Les coefficients obtenus dépendent de la précision des mesures, mais l'attracteur, lui, dépend continûment des coefficients : de petites variations des coefficients modifient peu la fougère obtenue. Comme w_1 n'est pas une similitude (elle est même non inversible), ni la dimension de similitude, ni la dimension généralisée de la fougère ne sont définies.

Exercice 3.13. Plusieurs réponses sont possibles ; en voici une.

a) Les figures semblables et le rapport de similitude (deuxième et troisième secondaire) ; les rapports k et k^2 pour les longueurs et les aires (troisième secondaire) ; les régularités et les suites, puis les fonctions exponentielles (quatrième et cinquième secondaire) ; les logarithmes, pour calculer une dimension de similitude (cinquième secondaire, section 3.3) ; les probabilités, pour le jeu du chaos.

b) On peut faire dresser un tableau des périmètres et des aires des premières étapes, puis demander aux élèves de prévoir la suite. Le périmètre est multiplié par $\frac{4}{3}$ à chaque étape et dépasse tout nombre fixé à l'avance ; l'aire, elle, reste toujours inférieure à l'aire d'un cercle qui contient la figure. Cet argument géométrique suffit à montrer que l'aire est bornée, sans calculer la somme d'une série. Il faut toutefois éviter de laisser croire que le flocon « s'arrête » à l'étape dessinée : les figures tracées sont des approximations de l'objet limite. Il faut aussi éviter de présenter le paradoxe comme une contradiction, car le périmètre et l'aire mesurent des grandeurs différentes.

c) Le jeu à la main, avec un dé et une règle, rend le procédé transparent : chaque élève comprend la règle et contribue quelques points. En regroupant les feuilles de la classe, on voit apparaître un début de structure. Il faut en revanche des milliers de points pour voir apparaître le triangle de Sierpiński, d'où l'intérêt de programmer ensuite la même règle (section 3.4). Le programme permet aussi de modifier facilement un paramètre (la fraction du chemin, le nombre de sommets). La limite principale est que l'apparition d'une figure déterministe à partir d'un procédé aléatoire paraît « magique » ; sans le théorème de Hutchinson, l'enseignant ne peut que l'expliquer partiellement, par exemple en remarquant que le point ne tombe jamais dans le triangle central.

Chapitre 4 – Les chaînes de Markov et l'algorithme PageRank de Google

Exercice 4.1. Comme $\mathbf{w} = (-5, 10, -30) = -5\,(1, -2, 6) = -5\,\mathbf{v}$, les deux vecteurs sont colinéaires. Or, dès que $\mathbf{v}$ est un vecteur propre associé à λ, tout multiple non nul de $\mathbf{v}$ en est un aussi (exercice 4.2) : $A\mathbf{w} = -5\,A\mathbf{v} = -5\lambda\mathbf{v} = \lambda\mathbf{w}$. Un vecteur propre n'est défini qu'à un facteur non nul près, et chaque logiciel choisit sa propre normalisation ; `numpy`, par exemple, renvoie des vecteurs propres de norme 1.

Exercice 4.2. Par linéarité du produit matriciel, $A(\mu\mathbf{v}) = \mu\,A\mathbf{v} = \mu\lambda\mathbf{v} = \lambda\,(\mu\mathbf{v})$. De plus, $\mu\mathbf{v} \neq \mathbf{0}$, puisque $\mu \neq 0$ et $\mathbf{v} \neq \mathbf{0}$; c'est donc un vecteur propre associé à λ. Pour $\mu = 0$, l'égalité $A\mathbf{0} = \lambda\mathbf{0}$ est vraie, mais le vecteur nul est exclu par la définition D.10 : il vérifierait l'égalité pour *toute* valeur de λ, et la notion de valeur propre perdrait son sens.

Exercice 4.3. Dans chaque cas, t désigne un nombre réel non nul.

a) $p(\lambda) = -\lambda^3 + 6\lambda^2 + 9\lambda - 54 = -(\lambda - 6)(\lambda - 3)(\lambda + 3)$. Les valeurs propres sont 6, 3 et -3, de vecteurs propres respectifs $(t, 0, 0)$, $(t, 3t, -5t)$ et $(t, 0, -t)$.

b) $p(\lambda) = -\lambda^3 + 7\lambda^2 + 8\lambda = -\lambda(\lambda - 8)(\lambda + 1)$. Les valeurs propres sont 8, -1 et 0, de vecteurs propres respectifs $(5t, 4t, 9t)$, $(t, -t, 0)$ et $(t, -2t, t)$.

c) $p(\lambda) = (a-\lambda)(b-\lambda)$. Si $a \neq b$, les valeurs propres sont a et b, de vecteurs propres respectifs $(t, 0)$ et $(0, t)$. Si $a = b$, la matrice vaut aI : la seule valeur propre est a, et tout vecteur non nul est un vecteur propre (c'est le cas, en particulier, de la matrice nulle, avec $a = b = 0$).

d) $p(\lambda) = -\lambda^3 + \frac{11}{6}\lambda^2 - \frac{5}{6}\lambda = -\frac{1}{6}\lambda(\lambda-1)(6\lambda-5)$. Les valeurs propres sont 1, $\frac{5}{6}$ et 0, de vecteurs propres respectifs $(t,0,0)$, $(-2t,t,t)$ et $(t,2t,-3t)$. La matrice est stochastique : ses coefficients sont positifs et chaque colonne a une somme égale à 1. L'état 1 est *absorbant* : une fois atteint, on y reste avec la probabilité 1. Le vecteur de probabilités $(1,0,0)$ est l'état d'équilibre, et la valeur propre $\frac{5}{6}$ indique que la probabilité de ne pas encore avoir été absorbé diminue d'un facteur $\frac{5}{6}$ à chaque étape.

Exercice 4.4. Montrons par récurrence que $P^n\mathbf{v} = \lambda^n\mathbf{v}$. Pour $n = 1$, c'est l'hypothèse. Supposons l'égalité vraie pour un certain $n \geq 1$. Alors

$$P^{n+1}\mathbf{v} = P\left(P^n\mathbf{v}\right) = P\left(\lambda^n\mathbf{v}\right) = \lambda^n\, P\mathbf{v} = \lambda^n\lambda\mathbf{v} = \lambda^{n+1}\mathbf{v},$$

ce qui établit l'égalité au rang $n+1$. Comme $\mathbf{v} \neq \mathbf{0}$, l'égalité $P^n\mathbf{v} = \lambda^n\mathbf{v}$ signifie que λ^n est une valeur propre de P^n, associée au vecteur propre $\mathbf{v}$. Ce résultat est utilisé dans la démonstration du théorème 4.6.

Exercice 4.5. Soient $\mathbf{v}$ et $\mathbf{w}$ deux vecteurs de probabilités colinéaires : puisque $\mathbf{w} \neq \mathbf{0}$, il existe un nombre c tel que $\mathbf{v} = c\,\mathbf{w}$. En additionnant les composantes, on obtient $1 = c \cdot 1$, donc $c = 1$ et $\mathbf{v} = \mathbf{w}$. Deux vecteurs de probabilités colinéaires sont donc égaux. Si 1 est une valeur propre simple, ses vecteurs propres forment une droite, et deux états d'équilibre seraient deux vecteurs de probabilités colinéaires, donc égaux.

Exercice 4.6. Non. Par exemple, $M = \begin{pmatrix} 1 & \frac{1}{2} \\ 0 & 2 \end{pmatrix}$ vérifie $M\begin{pmatrix}1\\0\end{pmatrix} = \begin{pmatrix}1\\0\end{pmatrix}$: le vecteur de probabilités $(1,0)$ est un vecteur propre associé à la valeur propre 1. Pourtant, M n'est pas stochastique, puisque la somme de sa seconde colonne vaut $\frac{5}{2}$. Le vecteur propre ne renseigne que sur l'action de M dans une seule direction.

Exercice 4.7. La i^{e} composante de $M\mathbf{v}$ est $\sum_j m_{ij}v_j$; c'est une somme de produits de nombres positifs, donc elle est positive. La somme des composantes vaut

$$\sum_{i=1}^{N}\sum_{j=1}^{N} m_{ij}v_j = \sum_{j=1}^{N} v_j\Big(\sum_{i=1}^{N} m_{ij}\Big) = \sum_{j=1}^{N} v_j \cdot 1 = 1,$$

puisque la somme de chaque colonne de M vaut 1 et que celle des composantes de $\mathbf{v}$ vaut 1. Par récurrence, $M^n\mathbf{v}$ est un vecteur de probabilités pour tout n.

Exercice 4.8.

a) C'est le calcul de la démonstration du théorème 4.8 : par l'inégalité triangulaire, $\|M\mathbf{v}\|_1 = \sum_i \left|\sum_j m_{ij}v_j\right| \leq \sum_i \sum_j m_{ij}|v_j| = \sum_j |v_j| \sum_i m_{ij} = \|\mathbf{v}\|_1$.

b) Si $M\mathbf{v} = \lambda\mathbf{v}$ avec $\mathbf{v} \neq \mathbf{0}$ réel, alors $|\lambda|\,\|\mathbf{v}\|_1 = \|M\mathbf{v}\|_1 \leq \|\mathbf{v}\|_1$, et l'on divise par $\|\mathbf{v}\|_1 > 0$.

c) Si $M\mathbf{v} = \lambda\mathbf{v}$ et que $\mathbf{v}$ est un vecteur de probabilités, la somme des composantes de $M\mathbf{v}$ vaut 1 (exercice 4.7), et celle de $\lambda\mathbf{v}$ vaut λ. Donc $\lambda = 1$.

Exercice 4.9. Un vecteur de probabilités colinéaire à $\mathbf{v}$ s'écrit $c\,\mathbf{v}$, et la condition sur la somme impose $c = 1/(v_1 + \cdots + v_N)$. Il existe si et seulement si $\mathbf{v} \neq \mathbf{0}$ et si toutes ses composantes sont de même signe (positives ou nulles, ou négatives ou nulles) ; il est alors unique (exercice 4.5).

```python
import numpy as np

def vers_probabilites(v, tol=1e-12):
    v = np.real(np.asarray(v, dtype=complex))
    s = v.sum()
    if abs(s) < tol:
        raise ValueError("la somme des composantes est nulle")
    p = v / s
    if (p < -tol).any():        # erreurs d'arrondi tolérées
        raise ValueError("composantes de signes différents")
    return np.clip(p, 0, None)  # remplace -0.0 par 0

M = np.array([[0.8, 0.4], [0.2, 0.6]])
valeurs, vecteurs = np.linalg.eig(M)
print(vers_probabilites(vecteurs[:, 0]))  # [0.66666667 0.33333333]
```

Le vecteur propre associé à 1 que renvoie `numpy` est $(0{,}894\,;\ 0{,}447)$, de norme 1 ; il est parfois renvoyé avec des composantes négatives, ce que la division par la somme corrige automatiquement. Le vecteur propre associé à $0{,}4$, proportionnel à $(1, -1)$, a une somme nulle : la fonction signale qu'aucun vecteur de probabilités ne lui est colinéaire.

Exercice 4.10.

a) On calcule $M_2\mathbf{q} = \frac{1}{32}(0{,}9\cdot 21 + 0{,}7\cdot 3,\ 0{,}5\cdot 3 + 0{,}3\cdot 5,\ 0{,}1\cdot 21 + 0{,}3\cdot 3,\ 0{,}5\cdot 3 + 0{,}7\cdot 5) = \frac{1}{32}(21, 3, 3, 5) = \mathbf{q}$, et la somme des composantes vaut 1. La matrice M_2 est régulière (tous les coefficients de M_2^2 sont strictement positifs), donc cet équilibre est unique.

b) On part de l'état PB. Demain, on passe à BB avec la probabilité $0{,}7$, puis il fait beau après-demain avec la probabilité $0{,}9$; ou l'on passe à BP avec la probabilité $0{,}3$, puis il fait beau avec la probabilité $0{,}5$. La probabilité cherchée est $0{,}7\cdot 0{,}9 + 0{,}3\cdot 0{,}5 = 0{,}78$.

c) C'est $q_{PP} = \frac{5}{32} \approx 0{,}156$. Dans le modèle à deux états, elle vaut $\mathbb{P}(X_n = P)\cdot 0{,}6 = \frac{1}{3}\cdot 0{,}6 = 0{,}2$. Le modèle à mémoire, où il fait aussi plus souvent beau à long terme ($\frac{3}{4}$ au lieu de $\frac{2}{3}$), produit moins de séquences de pluie.

Exercice 4.11. Dans l'ordre A, V, C, la matrice de transition et la répartition initiale sont

$$M = \begin{pmatrix} 0{,}96 & 0{,}02 & 0{,}02 \\ 0{,}03 & 0{,}95 & 0{,}05 \\ 0{,}01 & 0{,}03 & 0{,}93 \end{pmatrix}, \qquad \mathbf{x}_0 = \begin{pmatrix} 0{,}65 \\ 0{,}30 \\ 0{,}05 \end{pmatrix}.$$

On obtient

$$\mathbf{x}_1 = \begin{pmatrix} 0{,}631 \\ 0{,}307 \\ 0{,}062 \end{pmatrix}, \quad \mathbf{x}_5 \approx \begin{pmatrix} 0{,}5657 \\ 0{,}3319 \\ 0{,}1024 \end{pmatrix}, \quad \mathbf{x}_{10} \approx \begin{pmatrix} 0{,}5039 \\ 0{,}3568 \\ 0{,}1393 \end{pmatrix}.$$

À long terme, la répartition se rapproche de l'état d'équilibre $\mathbf{q} = (\frac{1}{3}, \frac{13}{30}, \frac{7}{30})$, soit environ $33{,}3\,\%$ de A, $43{,}3\,\%$ de V et $23{,}3\,\%$ de C. Les valeurs propres de M sont 1, $0{,}94$ et $0{,}9$: la convergence est lente, et la répartition est encore loin de l'équilibre après dix ans, puisque $(0{,}94)^{10} \approx 0{,}54$.

Exercice 4.12. Dans l'ordre dormir (D), manger (M), courir (R), la matrice de transition

est

$$H = \begin{pmatrix} 0{,}9 & 0{,}7 & 0{,}8 \\ 0{,}05 & 0 & 0 \\ 0{,}05 & 0{,}3 & 0{,}2 \end{pmatrix}.$$

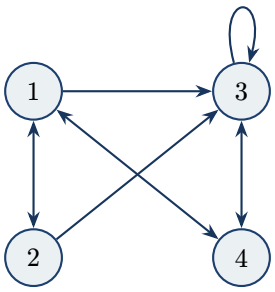
Le système $(H - I)\mathbf{q} = \mathbf{0}$ donne $q_M = 0{,}05\, q_D$, puis, par la troisième ligne, $0{,}8\, q_R = 0{,}05\, q_D + 0{,}3\, q_M = 0{,}065\, q_D$, soit $q_R = \frac{13}{160}\, q_D$. Avec $q_D + q_M + q_R = 1$, on trouve $\mathbf{q} = \frac{1}{181}(160, 8, 13) \approx (0{,}884\,;\ 0{,}044\,;\ 0{,}072)$. La matrice H est régulière (H^2 n'a que des coefficients strictement positifs) ; par le théorème ergodique (remarque 4.2), Doudou passe environ $88{,}4\,\%$ de son temps à dormir, soit environ $0{,}884 \times 24 \approx 21{,}2$ heures par jour, ou 21 h 13 min.

Exercice 4.13. Les voisines de chaque intersection sont : $1 : 2, 4$; $2 : 1, 5$; $3 : 4, 6$; $4 : 1, 3, 5, 7$; $5 : 2, 4, 8$; $6 : 3, 7$; $7 : 4, 6, 8$; $8 : 5, 7$. Depuis une intersection qui a d voisines, le policier se rend à chacune des $d + 1$ possibilités (en comptant l'intersection elle-même) avec la probabilité $1/(d + 1)$. On construit la matrice de transition 8×8 avec la fonction `matrice_transition` du programme 4.5, en ajoutant chaque intersection à la liste de ses voisines, puis on calcule l'état d'équilibre :

$$\mathbf{q} = \tfrac{1}{28}\,(3, 3, 3, 5, 4, 3, 4, 3).$$

Le policier passe donc environ $10{,}7\,\%$ de son temps à chacune des intersections 1, 2, 3, 6 et 8, environ $17{,}9\,\%$ à l'intersection 4, et environ $14{,}3\,\%$ à chacune des intersections 5 et 7. On remarque que q_i est proportionnel à $d_i + 1$. Ce n'est pas un hasard : si $q_j = c\,(d_j + 1)$ pour tout j, la i^{e} composante de $M\mathbf{q}$ est la somme, sur les $d_i + 1$ intersections j voisines de i ou égales à i, des termes $\frac{1}{d_j+1}\, c\,(d_j + 1) = c$, soit $c\,(d_i + 1) = q_i$. La relation de voisinage étant symétrique, chaque intersection reçoit autant qu'elle donne.

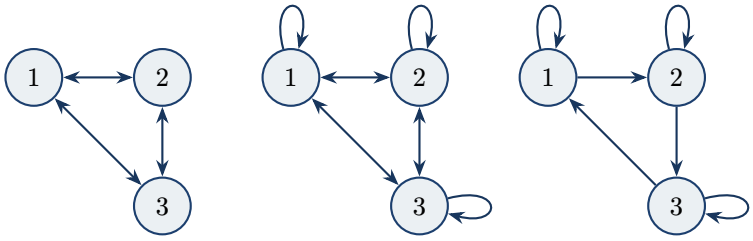
Exercice 4.14. La colonne j décrit les liens qui partent de la page j : la page 1 pointe vers 2, 3 et 4 ; la page 2, vers 1 et 3 ; la page 3, vers elle-même et vers 4 ; la page 4, vers 1 et 3. Le coefficient $a_{33} = \frac{1}{2}$ correspond à un lien de la page 3 vers elle-même.



Exercice 4.15.

a) L'internaute parcourt les trois pages à tour de rôle, et passe autant de temps sur chacune : le PageRank est $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. On peut aussi invoquer la symétrie : la permutation $1 \to 2 \to 3 \to 1$ conserve le mini-web, donc les trois pages jouent le même rôle.

b) Oui. Il suffit d'ajouter des liens qui préservent cette symétrie, par exemple tous les liens dans les deux sens, une boucle sur chaque page, ou les deux.



Plus généralement, le PageRank reste uniforme dès que la somme de chaque *ligne* de la matrice de transition vaut aussi 1 : alors $A\mathbf{e} = \mathbf{e}$, et la loi uniforme est un état d'équilibre.

c) Non : partant de la page 1, l'internaute est sur la page 1 aux étapes $0, 3, 6, \ldots$, et la loi $\mathbf{p}_n$ parcourt cycliquement $(1,0,0)$, $(0,1,0)$, $(0,0,1)$. La chaîne est périodique, comme celle de l'exemple 4.5, et la matrice n'est pas régulière. Le PageRank s'interprète alors comme la proportion du temps passé sur chaque page (remarque 4.2), qui vaut bien $\frac{1}{3}$. Dans la version de la troisième figure, les boucles rendent la matrice régulière, et $\mathbf{p}_n$ converge.

Exercice 4.16.

a) De la page 5, l'internaute passe à la page 3, puis il fait indéfiniment l'aller-retour entre les pages 3 et 1 : à long terme, il passe la moitié de son temps sur chacune des pages 1 et 3, et aucun temps sur les autres.

b) La page 2 n'a aucun lien sortant : la promenade s'arrête, et le modèle n'est pas défini. C'est le problème de la section 4.5.2.

c) La matrice de transition est

$$M = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{3} & 0 \\ 1 & 0 & 0 & \frac{1}{3} & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{3} & 0 \end{pmatrix}.$$

Sa deuxième colonne est nulle : M n'est pas stochastique.

d) $\mathbf{v} = (1,0,1,0,0)$. Le noyau de $M - I$ est la droite engendrée par $\mathbf{v}$; le vecteur de probabilités correspondant, $(\frac{1}{2}, 0, \frac{1}{2}, 0, 0)$, retrouve la réponse de a).

e) $p(\lambda) = \det(M - \lambda I) = -\lambda^5 + \lambda^3 = -\lambda^3(\lambda - 1)(\lambda + 1)$. Les valeurs propres sont 0 (de multiplicité 3), 1 et -1 ; la valeur propre -1 traduit l'aller-retour entre les pages 1 et 3.

f) On remplace la deuxième colonne par une colonne de $\frac{1}{5}$: $N = M + \frac{1}{5}\,\mathbf{e}\,\mathbf{E}^\intercal$, avec $\mathbf{E} = (0,1,0,0,0)$.

g) Avec $\beta = 0{,}9$, $G = 0{,}9\,N + 0{,}1 \cdot \frac{1}{5}\,\mathbf{e}\,\mathbf{e}^\intercal$:

$$G = \begin{pmatrix} \frac{1}{50} & \frac{1}{5} & \frac{23}{25} & \frac{1}{50} & \frac{1}{50} \\ \frac{1}{50} & \frac{1}{5} & \frac{1}{50} & \frac{8}{25} & \frac{1}{50} \\ \frac{23}{25} & \frac{1}{5} & \frac{1}{50} & \frac{8}{25} & \frac{23}{25} \\ \frac{1}{50} & \frac{1}{5} & \frac{1}{50} & \frac{1}{50} & \frac{1}{50} \\ \frac{1}{50} & \frac{1}{5} & \frac{1}{50} & \frac{8}{25} & \frac{1}{50} \end{pmatrix}.$$

Son état d'équilibre est le PageRank $\mathbf{q} \approx (0{,}4429\,;\ 0{,}0339\,;\ 0{,}4631\,;\ 0{,}0261\,;\ 0{,}0339)$. Le classement est : page 3, page 1, pages 2 et 5 à égalité, page 4. La page 4, vers laquelle ne pointe aucun lien, n'est visitée que par téléportation.

Exercice 4.17.

a) Les liens sortants sont : $1 \to 2$; $2 \to 1,3$; $3 \to 1,2,5$; $4 \to 1,6$; $5 \to 3,4$; $6 \to 7$; $7 \to 6$. D'où

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

Chaque page a au moins un lien sortant, mais les pages 6 et 7 forment une région piégeante : sans téléportation, l'unique état d'équilibre est $\left(0,0,0,0,0,\frac{1}{2},\frac{1}{2}\right)$.

b) Puisque $\mathbf{E} = \mathbf{0}$, on a $B = A$ et $G = 0{,}85\,A + 0{,}15 \cdot \frac{1}{7}\,\mathbf{e}\,\mathbf{e}^\intercal$: chaque coefficient de G vaut $0{,}85\,a_{ij} + \frac{0{,}15}{7}$.

c) Avec la fonction `pagerank` du programme 4.9, ou la fonction `equilibre` du programme 4.7, on obtient

$$\mathbf{q} \approx (0{,}1593\,;\ 0{,}1931\,;\ 0{,}1280\,;\ 0{,}0460\,;\ 0{,}0577\,;\ 0{,}2132\,;\ 0{,}2027).$$

Le classement est : page 6, page 7, page 2, page 1, page 3, page 5, page 4. Comme pour le mini-web 3, la région piégeante occupe les deux premières places, avec environ 42 % du PageRank total.

Exercice 4.18.

a) Le coefficient (i,j) de $\mathbf{e}\,\mathbf{E}^\intercal$ vaut $e_i E_j = E_j$. La matrice $\frac{1}{N}\,\mathbf{e}\,\mathbf{E}^\intercal$ a donc pour j^{e} colonne le vecteur $\frac{1}{N}\,\mathbf{e}$, dont toutes les composantes valent $\frac{1}{N}$, si la page j n'a pas de lien sortant, et le vecteur nul sinon. C'est la loi uniforme, placée dans les colonnes des pages sans liens sortants.

b) $B = A + \frac{1}{N}\,\mathbf{e}\,\mathbf{E}^\intercal$: on ajoute la loi uniforme dans les colonnes nulles de A, et dans celles-là seulement.

c) $G = \beta B + (1-\beta)\,\frac{1}{N}\,\mathbf{e}\,\mathbf{e}^\intercal = \beta A + \frac{\beta}{N}\,\mathbf{e}\,\mathbf{E}^\intercal + \frac{1-\beta}{N}\,\mathbf{e}\,\mathbf{e}^\intercal$, et l'on met $\frac{1}{N}\,\mathbf{e}$ en facteur à gauche dans les deux derniers termes.

d) La j^{e} colonne de A a autant de coefficients non nuls que la page j a de liens sortants, soit quelques dizaines au plus, alors que N se compte en milliards : la proportion de coefficients non nuls est infime. Pour G, chaque coefficient est au moins égal à $(1-\beta)/N > 0$. Comme la somme de chaque colonne de G vaut 1 et que chaque colonne contient $N \geq 2$ coefficients strictement positifs, chacun d'eux est strictement inférieur à 1.

Exercice 4.19. Plusieurs réponses sont possibles ; en voici une.

a) Les *probabilités conditionnelles* et le *diagramme en arbre* : la matrice de transition regroupe des probabilités conditionnelles, et la formule $\mathbf{x}_{n+1} = M\mathbf{x}_n$ est la règle « multiplier le long des branches, additionner les branches » (remarque 4.1). Les *graphes orientés valués* : le diagramme de transition et le mini-web en sont (remarque 4.2). La *multiplication matricielle*, dans la séquence Technico-sciences : elle permet de calculer la loi d'un jour à partir de celle de la veille. On pourrait ajouter l'espérance mathématique (exemple 4.2) et la comparaison entre probabilités fréquentielles et théoriques (remarque 4.3).

b) Les élèves reçoivent le diagramme du modèle météorologique et un dé à dix faces. En équipe, ils simulent trente jours, d'abord en partant d'un jour de beau temps, puis d'un jour de pluie. On compile au tableau la proportion de beaux jours pour chaque équipe, et l'on compare les deux points de départ : la proportion est proche de $\frac{2}{3}$ dans les deux cas. On calcule ensuite, à l'aide d'un arbre, la probabilité qu'il fasse beau dans deux jours, puis l'on montre qu'un tableur permet de prolonger le calcul. L'activité permet d'aborder l'idée que le hasard « n'a pas de mémoire », souvent étendue à tort à toutes les situations : ici, au contraire, le temps d'un jour dépend de celui de la veille, et les jours de pluie se suivent en séquences. Elle permet aussi de discuter du biais d'équiprobabilité : il n'y a que deux états, mais ils ne sont pas également probables.

c) Les élèves consultent quotidiennement des classements produits par des algorithmes : résultats de recherche, fils d'actualité, suggestions de vidéos. Étudier PageRank leur montre qu'un tel classement repose sur un *modèle*, fondé sur des choix qui ne sont pas neutres : la popularité n'est pas la fiabilité, le facteur β est choisi par le concepteur, et des régions piégeantes peuvent être créées pour manipuler le classement. Comprendre qu'une page bien classée n'est pas nécessairement une page exacte est une compétence citoyenne.

Chapitre 5 – L'analyse en composantes principales

Exercice 5.1.

a) $M = (3, 4)$; les points centrés sont $(-2, -3)$, $(0, -1)$, $(0, 3)$ et $(2, 1)$.

b) $s_x^2 = \frac{4+0+0+4}{4} = 2$, $s_y^2 = \frac{9+1+9+1}{4} = 5$ et $s_{xy} = \frac{6+0+0+2}{4} = 2$, d'où $C = \begin{pmatrix} 2 & 2 \\ 2 & 5 \end{pmatrix}$ et $r_{xy} = \frac{2}{\sqrt{10}} \approx 0{,}632$.

c) $\det(C - \lambda I) = \lambda^2 - 7\lambda + 6 = (\lambda - 6)(\lambda - 1)$. Pour $\lambda_1 = 6$, $C - 6I = \begin{pmatrix} -4 & 2 \\ 2 & -1 \end{pmatrix}$ a pour noyau $\operatorname{span}\{(1, 2)\}$; pour $\lambda_2 = 1$, $C - I = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$ a pour noyau $\operatorname{span}\{(2, -1)\}$. Vecteurs unitaires : $\mathbf{v}_1 = \frac{1}{\sqrt{5}}(1, 2)$ et $\mathbf{v}_2 = \frac{1}{\sqrt{5}}(2, -1)$, qui sont bien orthogonaux.

d) L'axe principal passe par $(3, 4)$ avec une pente de 2 : $y = 2x - 2$. La droite de régression a pour pente $s_{xy}/s_x^2 = 1$ et passe aussi par M : $y = x + 1$. L'axe principal est plus pentu, conformément à la remarque 5.3.

e) $t_i = \frac{1}{\sqrt{5}}(\Delta x_i + 2\,\Delta y_i)$, d'où $\sqrt{5}\,(t_1, t_2, t_3, t_4) = (-8, -2, 6, 4)$. La moyenne des t_i est nulle et $\frac{1}{4}\sum t_i^2 = \frac{64+4+36+16}{4 \cdot 5} = \frac{120}{20} = 6 = \lambda_1$.

f) $\frac{6}{6+1} \approx 85{,}7\,\%$. La moyenne des carrés des distances vaut $\lambda_2 = 1$, donc leur somme vaut 4. On le vérifie : les distances signées sont $\frac{1}{\sqrt{5}}(2\,\Delta x_i - \Delta y_i)$, soit $\frac{1}{\sqrt{5}}(-1, 1, -3, 3)$, et $\frac{1+1+9+9}{5} = 4$.

Exercice 5.2.

a) $f(\theta) = \mathbf{u}^{\mathsf{T}} C\,\mathbf{u} = 13\cos^2\theta + 6\sqrt{3}\sin\theta\cos\theta + 7\sin^2\theta$. Avec $\cos^2\theta = \frac{1+\cos 2\theta}{2}$, $\sin^2\theta = \frac{1-\cos 2\theta}{2}$ et $2\sin\theta\cos\theta = \sin 2\theta$, on obtient $f(\theta) = 10 + 3\cos 2\theta + 3\sqrt{3}\sin 2\theta$.

b) On cherche $A > 0$ et un angle φ tels que $3\cos 2\theta + 3\sqrt{3}\sin 2\theta = A\cos(2\theta - \varphi)$ pour tout θ. Par la formule d'addition, $A\cos(2\theta - \varphi) = (A\cos\varphi)\cos 2\theta +$

$(A\sin\varphi)\sin 2\theta$; il suffit donc que

$$A\cos\varphi = 3 \qquad \text{et} \qquad A\sin\varphi = 3\sqrt{3}.$$

En élevant ces deux égalités au carré et en les additionnant, on obtient $A^2(\cos^2\varphi + \sin^2\varphi) = 3^2 + (3\sqrt{3})^2$, c'est-à-dire $A^2 = 9 + 27 = 36$, d'où $A = 6$. On en tire $\cos\varphi = \frac{3}{6} = \frac{1}{2}$ et $\sin\varphi = \frac{3\sqrt{3}}{6} = \frac{\sqrt{3}}{2}$, donc $\varphi = 60° : f(\theta) = 10 + 6\cos(2\theta - 60°)$. Le maximum vaut 16, pour $2\theta = 60°$, soit $\theta = 30°$; le minimum vaut 4, pour $\theta = 120°$.

c) $\det(C - \lambda I) = \lambda^2 - 20\lambda + (91 - 27) = \lambda^2 - 20\lambda + 64 = (\lambda - 16)(\lambda - 4)$. On vérifie que $C\begin{pmatrix}\sqrt{3}/2\\ 1/2\end{pmatrix} = \begin{pmatrix}13\sqrt{3}/2 + 3\sqrt{3}/2\\ 9/2 + 7/2\end{pmatrix} = 16\begin{pmatrix}\sqrt{3}/2\\ 1/2\end{pmatrix}$: le vecteur propre associé à 16 est bien $(\cos 30°, \sin 30°)$, et celui associé à 4 lui est perpendiculaire, d'angle 120°.

Exercice 5.3.

a) En développant selon la troisième ligne, $\det(R - \lambda I) = (1-\lambda)\big((1-\lambda)^2 - 0{,}64\big)$, dont les racines sont $\lambda_1 = 1{,}8$, $\lambda_2 = 1$ et $\lambda_3 = 0{,}2$. Les vecteurs propres unitaires sont $\mathbf{v}_1 = \frac{1}{\sqrt{2}}(1, 1, 0)$, $\mathbf{v}_2 = (0, 0, 1)$ et $\mathbf{v}_3 = \frac{1}{\sqrt{2}}(1, -1, 0)$. La somme $1{,}8 + 1 + 0{,}2 = 3$ est bien la trace de R, c'est-à-dire la somme des variances des trois variables réduites.

b) 60 %, 33,3 % et 6,7 %. Il faut deux axes, qui expliquent 93,3 % de la variance. Le critère de Kaiser retient le premier axe ; le deuxième, dont la valeur propre vaut exactement 1, est à la limite : il résume autant d'information qu'une variable de départ, et pour cause, puisqu'il *est* la variable « éducation physique ».

c) Le premier axe, proportionnel à $z_{\text{math}} + z_{\text{phys}}$, mesure la réussite globale dans les deux matières scientifiques. Le deuxième est l'éducation physique, non corrélée aux deux autres. Le troisième, proportionnel à $z_{\text{math}} - z_{\text{phys}}$, oppose les élèves relativement plus forts en mathématiques à ceux qui le sont en physique ; sa faible variance traduit la forte corrélation entre ces deux matières.

d) $y_1 = \frac{1{,}5+0{,}5}{\sqrt{2}} = \sqrt{2} \approx 1{,}414$ et $y_2 = -1$. La troisième coordonnée est $y_3 = \frac{1{,}5-0{,}5}{\sqrt{2}} = \frac{1}{\sqrt{2}}$, et le carré de la distance au plan vaut $y_3^2 = 0{,}5$. On vérifie que $y_1^2 + y_2^2 + y_3^2 = 2 + 1 + 0{,}5 = 3{,}5 = 1{,}5^2 + 0{,}5^2 + 1^2$.

e) La reconstruction est $y_1\mathbf{v}_1 + y_2\mathbf{v}_2 = (1, 1, 0) + (0, 0, -1) = (1, 1, -1)$. L'écart avec les cotes de départ est $(0{,}5\,;\ -0{,}5\,;\ 0)$, dont le carré de la norme vaut 0,5 : le carré de la distance au plan est l'erreur quadratique de reconstruction de cet élève. D'après le théorème 5.4 (point 4), la moyenne de cette erreur sur tous les élèves vaut $\lambda_3 = 0{,}2$. Cet élève est donc moins bien représenté que la moyenne par le plan principal. C'est logique : il est nettement plus fort en mathématiques qu'en physique, et cet écart est justement ce que mesure le troisième axe, celui que l'on abandonne.

Exercice 5.4.

a) $s_{y'}^2 = \frac{1}{100}\,s_y^2 = 0{,}5$ et $s_{xy'} = \frac{1}{10}\,s_{xy} = 4$, d'où $C' = \begin{pmatrix}50 & 4\\ 4 & 0{,}5\end{pmatrix}$. Le coefficient de corrélation est inchangé : $\frac{4}{\sqrt{50}\sqrt{0{,}5}} = \frac{4}{5} = 0{,}8$.

b) Le polynôme caractéristique est $\lambda^2 - 50{,}5\,\lambda + 9$, de racines $\lambda = \frac{50{,}5 \pm \sqrt{2550{,}25 - 36}}{2}$, soit $\lambda_1 \approx 50{,}32$ et $\lambda_2 \approx 0{,}18$. Un vecteur propre associé à λ_1 est $(4\,;\ \lambda_1 - 50) \approx (4\,;\ 0{,}32)$, c'est-à-dire une pente d'environ 0,080. Le premier axe explique $\frac{50{,}32}{50{,}5} \approx 99{,}6\,\%$ de la variance.

c) Une pente de 0,080 en points sur 10 par point sur 100 correspond à une pente de 0,80 dans les unités de départ : l'axe principal est devenu, à très peu près, la droite de régression $y = 0{,}8\,x + 19$, et non plus $y = x + 5$. En divisant la variance des sciences par 100, on a presque fait disparaître cette variable de la variance totale, et le premier axe se contente de suivre la direction des mathématiques. Les résultats d'une ACP non réduite dépendent donc des unités de mesure.

d) La cote z d'une valeur ne change pas si l'on multiplie la variable par une constante positive c : $\frac{cx_i - c\bar{x}}{c\,s_x} = \frac{x_i - \bar{x}}{s_x}$. Le tableau réduit, donc la matrice de corrélation et toute l'ACP, sont inchangés.

Exercice 5.5.

a) Les écarts de G et de H au point $(70, 75)$ sont $(k, -k)$ et $(-k, k)$, de somme nulle : le point moyen ne change pas. Les sommes des carrés et des produits des écarts deviennent $300 + 2k^2$ (pour chaque variable) et $240 - 2k^2$, d'où

$$C_k = \frac{1}{8}\begin{pmatrix} 300 + 2k^2 & 240 - 2k^2 \\ 240 - 2k^2 & 300 + 2k^2 \end{pmatrix}.$$

b) Une matrice de la forme $\begin{pmatrix} \alpha & \beta \\ \beta & \alpha \end{pmatrix}$ vérifie $C(1,1) = (\alpha + \beta)(1,1)$ et $C(1,-1) = (\alpha - \beta)(1,-1)$. Les vecteurs propres sont donc $\mathbf{u}$ et $\mathbf{w}$, de valeurs propres $\lambda_{\mathbf{u}} = \frac{540}{8} = 67{,}5$, qui ne dépend pas de k, et $\lambda_{\mathbf{w}} = \frac{60 + 4k^2}{8} = 7{,}5 + \frac{k^2}{2}$. Les élèves G et H se projettent au point moyen sur l'axe dirigé par $\mathbf{u}$, mais très loin sur l'axe dirigé par $\mathbf{w}$.

c) Pour $k = 12$, $G = (82, 63)$ et $H = (58, 87)$. On a $s_x^2 = s_y^2 = \frac{588}{8} = 73{,}5$ et $s_{xy} = \frac{240 - 288}{8} = -6$, d'où $r_{xy} = \frac{-6}{73{,}5} \approx -0{,}08$. Les valeurs propres sont $\lambda_{\mathbf{w}} = 79{,}5$ et $\lambda_{\mathbf{u}} = 67{,}5$: l'axe principal est maintenant dirigé par $\mathbf{w}$, et il n'explique que $\frac{79{,}5}{147} \approx 54\,\%$ de la variance. Pour les six élèves, la corrélation valait 0,8, et l'axe dirigé par $\mathbf{u}$ expliquait $90\,\%$ de la variance. Deux élèves sur huit ont suffi à effacer la tendance commune aux six autres.

d) L'axe bascule lorsque $7{,}5 + \frac{k^2}{2} > 67{,}5$, c'est-à-dire lorsque $k^2 > 120$, soit $k > \sqrt{120} \approx 10{,}95$; c'est exactement le moment où s_{xy} devient négative. Comme les variances et les covariances sont des moyennes de carrés et de produits d'écarts, un point très éloigné du centre pèse beaucoup plus lourd que les autres : quelques valeurs extrêmes peuvent modifier les axes à elles seules. Avant d'interpréter une ACP, il faut donc regarder le nuage et repérer les individus atypiques, puis se demander s'il s'agit d'erreurs ou d'observations réelles.

Exercice 5.6.

a) Faux. C'est en général une combinaison linéaire de toutes les variables, comme $y_1 = \frac{1}{\sqrt{2}}(x + y)$ pour les six élèves.

b) Vrai. La matrice de covariance est alors diagonale ; ses vecteurs propres sont les vecteurs de la base standard, et ses valeurs propres sont les variances.

c) Faux. L'axe principal est situé entre les deux droites de régression (remarque 5.3) ; il ne coïncide avec elles que si les points sont alignés.

d) Vrai. C'est la trace de la matrice de corrélation, dont les p coefficients diagonaux valent 1.

e) Faux. La matrice de covariance ne retient que les moyennes des produits des écarts : un nuage en forme d'ellipse et un nuage formé de deux groupes séparés peuvent avoir la même matrice de covariance, et donc les mêmes axes principaux. Il faut toujours regarder le nuage projeté, et pas seulement les valeurs propres.

f) Faux. La proportion de 95 % est une *moyenne* sur tous les individus. La moyenne des carrés des distances au plan est petite, mais un individu particulier peut être loin du plan, et deux individus qui se projettent au même endroit peuvent être éloignés dans la direction des autres axes. On peut contrôler ce risque en calculant, pour chaque individu, la distance entre ce point et sa projection.

Exercice 5.7. Le programme suivant répond aux quatre questions.

```
import numpy as np
import matplotlib.pyplot as plt

x = np.array([60, 63, 67, 75, 76, 79], dtype=float)
y = np.array([72, 66, 67, 84, 82, 79], dtype=float)
X = np.column_stack([x, y])
Xc = X - X.mean(axis=0)

C = np.cov(X.T, bias=True)                  # [[50, 40], [40, 50]]
valeurs, vecteurs = np.linalg.eigh(C)       # [10, 90]
print(C, valeurs, vecteurs, sep="\n")

theta = np.linspace(0, np.pi, 181)
U = np.column_stack([np.cos(theta), np.sin(theta)])
f = ((Xc @ U.T) ** 2).mean(axis=0)          # une variance par angle
print("maximum :", f.max(), "pour", np.degrees(theta[f.argmax()]))

a, b = np.polyfit(x, y, 1)                  # pente 0.8, ordonnée 19
print(a, C[0, 1] / C[0, 0])
plt.scatter(x, y)
t = np.linspace(55, 85, 2)
plt.plot(t, t + 5, label="axe principal")
plt.plot(t, a * t + b, "--", label="régression de y selon x")
plt.plot(t, 1.25 * t - 12.5, "-.", label="régression de x selon y")
plt.axis("equal"); plt.legend(); plt.show()
```

a) On obtient $C = \begin{pmatrix} 50 & 40 \\ 40 & 50 \end{pmatrix}$, de valeurs propres 10 et 90, comme à la section 5.2.

b) Le maximum, 90, est atteint pour $\theta = 45°$; la courbe obtenue est celle de $f(\theta) = 50 + 40 \sin 2\theta$.

c) `np.polyfit` renvoie la pente 0,8 et l'ordonnée à l'origine 19 ; on a bien $s_{xy}/s_x^2 = 40/50$.

d) En échangeant x et y, les deux droites de régression échangent leurs rôles : la droite « de y selon x » devient la droite $x = 0{,}8\,y + 10$. L'axe principal, lui, ne change pas (à la symétrie près des axes du dessin) : il ne privilégie aucune des deux variables.

Exercice 5.8. Le programme suivant trace les figures et effectue les calculs.

```
import matplotlib.pyplot as plt

especes = manchots["species"].to_numpy()
couleurs = {"Adelie": "tab:red", "Chinstrap": "tab:blue",
            "Gentoo": "tab:orange"}

def dessiner(P, titres):
    for e, c in couleurs.items():
```

```
        plt.scatter(P[especes == e, 0], P[especes == e, 1],
                    s=8, color=c, label=e)
    plt.xlabel(titres[0]); plt.ylabel(titres[1]); plt.legend()
    plt.show()

dessiner(X[:, [0, 1]], ["longueur du bec", "hauteur du bec"])  # etc.
valeurs, vecteurs, Y = acp(X)
dessiner(Y[:, [0, 1]], ["y1", "y2"])
print(acp(X, reduire=False)[:2])
print(np.cov(Y.T, bias=True).round(6))
```

a) On obtient les trois nuages de la figure 5.7.

b) Avec les vecteurs propres renvoyés par `eigh`, les papous sont à gauche : le premier vecteur propre est l'opposé de celui de la section 5.3 (remarque 5.5). Le dessin est simplement retourné.

c) Sans réduction, les valeurs propres sont environ 212,3, 16,1 et 2,4 ; le premier axe explique 92,0 % de la variance, et son vecteur propre vaut environ $\pm(0{,}27\,;\ -0{,}08\,;\ 0{,}96)$: il est presque dirigé selon la nageoire. Comme expliqué à la section 5.3, la nageoire, de loin la plus dispersée, accapare la variance totale. Le pourcentage élevé est trompeur : il traduit une différence d'échelle, pas une structure.

d) La matrice de covariance de Y est diagonale, de diagonale $(2{,}003\,;\ 0{,}767\,;\ 0{,}230)$.

Exercice 5.9.

a) Le premier axe explique 99,99 % de la variance, et son vecteur propre est presque exactement $(0, 0, 0, 1)$: la masse, dont l'écart type est d'environ 800 g, écrase tout le reste. L'ACP non réduite n'a ici aucun sens.

b) Les valeurs propres sont 2,754, 0,773, 0,365 et 0,109, soit 68,8 %, 19,3 %, 9,1 % et 2,7 %. Le critère de Kaiser ne retient qu'un axe ; il en faut deux pour dépasser 80 % (88,2 %). Les critères ne sont pas toujours d'accord.

c) Au signe près, $\mathbf{v}_1 \approx (0{,}46\,;\ -0{,}40\,;\ 0{,}58\,;\ 0{,}55)$: le premier axe oppose les gros manchots, à longue nageoire, long bec et bec mince (les papous), aux autres. C'est un axe de « taille ».

d) Le programme suivant effectue la vérification.

```
X4 = manchots[["bill_length_mm", "bill_depth_mm",
               "flipper_length_mm", "body_mass_g"]].to_numpy()
valeurs, V, Y = acp(X4)
Z = (X4 - X4.mean(axis=0)) / X4.std(axis=0)
Z_chapeau = Y[:, :2] @ V[:, :2].T
print(((Z - Z_chapeau) ** 2).sum(axis=1).mean(), valeurs[2:].sum()
    )
```

Les deux nombres valent 0,4737.

Exercice 5.10. Le programme suivant effectue les calculs.

```
X4e = X4.copy()
X4e[18, 3] = 42000                        # erreur de saisie
print(X4[:, 3].std(), X4e[:, 3].std())
print(np.corrcoef(X4[:, 2], X4[:, 3])[0, 1],
      np.corrcoef(X4e[:, 2], X4e[:, 3])[0, 1])
valeurs, V, Y = acp(X4e)
```

```
print(valeurs, valeurs / valeurs.sum())
print(V[:, :2])
print(Y[18, :2], np.abs(np.delete(Y, 18, axis=0)[:, :2]).max(axis=0))
Z = (X4e - X4e.mean(axis=0)) / X4e.std(axis=0)
print(np.abs(Z).max(axis=0))
```

a) L'écart type de la masse passe d'environ 801 g à environ 2192 g, presque trois fois plus, à cause d'un seul manchot sur 342. La corrélation entre la masse et la nageoire chute de 0,871 à 0,293.

b) Les valeurs propres deviennent 2,117, 0,961, 0,700 et 0,222, soit 52,9 %, 24,0 %, 17,5 % et 5,6 % ; le plan principal ne conserve plus que 77,0 % de la variance, au lieu de 88,2 %. Au signe près, le premier vecteur propre devient $(0{,}54\,;\ -0{,}46\,;\ 0{,}64\,;\ 0{,}31)$: la masse, qui y avait un poids comparable à celui de la nageoire, n'y joue plus qu'un rôle secondaire, et l'axe de « taille » est affaibli. Le deuxième vecteur propre, $\pm(0{,}14\,;\ 0{,}56\,;\ -0{,}10\,;\ 0{,}81)$, est maintenant dominé par la masse : un axe entier sert en bonne partie à décrire une seule valeur erronée.

c) Sur le deuxième axe, ce manchot a une coordonnée d'environ 15,3 en valeur absolue, alors que celles de tous les autres manchots restent inférieures à 1,32. Il est isolé, très loin du reste du nuage.

d) La cote z de la masse erronée vaut environ 17,2, alors qu'aucune autre mesure ne dépasse 3 en valeur absolue. Un simple examen des valeurs extrêmes de chaque variable (avec `X4e.max(axis=0)`, un histogramme ou les cotes z) aurait révélé l'erreur : aucun manchot ne pèse 42 kg. La réduction ne protège pas, car elle divise la masse par un écart type lui-même gonflé par l'erreur ; elle met les variables à la même échelle, mais ne change pas la position relative d'un individu aberrant, qui reste à 17 écarts types du centre.

Exercice 5.11.

a) Les écarts types vont d'environ 0,12 (pour les phénols non flavonoïdes) à plus de 300 (pour la proline) : sans réduction, la proline expliquerait à elle seule 99,8 % de la variance.

b) Les valeurs propres commencent par 4,71, 2,50, 1,45, 0,92, 0,85 ; le plan principal conserve 55,4 % de la variance (sur un total de 13). Le critère de Kaiser retient trois axes, et il en faut cinq pour dépasser 80 %.

c) Les trois cépages forment trois groupes presque disjoints : le premier axe sépare les cépages 0 et 2, et le deuxième isole le cépage 1. Avec seulement la moitié de la variance, le plan principal suffit à révéler la structure, invisible dans $\mathbb{R}^{13}$.

d) On applique la classe `PCA` aux données réduites.

```
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
pca = PCA().fit(StandardScaler().fit_transform(X))
print(pca.explained_variance_ratio_[:3])   # [0.362 0.192 0.111]
print(pca.explained_variance_[:3])         # [4.732 2.511 1.454]
```

Les proportions sont identiques, mais les valeurs propres sont légèrement plus grandes, dans le rapport $\frac{178}{177}$: `scikit-learn` divise par $n - 1$ (remarque 5.2). Les vecteurs propres (`pca.components_`) sont les mêmes, au signe près.

Exercice 5.12.

Plusieurs réponses sont possibles ; en voici une.

a) Le *nuage de points* et la *corrélation* (deuxième cycle) : l'ACP part d'un nuage, et son premier axe décrit sa direction d'allongement, d'autant plus marquée que la corrélation est forte (remarque 5.1). La *droite de régression* : elle sert de point de comparaison avec l'axe principal, et la comparaison oblige à préciser ce que l'on minimise (remarque 5.3). Les *fonctions sinusoïdales* (cinquième secondaire, SN) : la variance dans une direction du plan est une fonction sinusoïdale de l'angle (remarque 5.2). On pourrait ajouter la *moyenne* et les *mesures de dispersion*, ainsi que les *vecteurs* et le *produit scalaire* (SN), qui servent à projeter.

b) On donne aux élèves, en équipe, une feuille de données décrivant une quarantaine de manchots par quatre mesures, sans l'espèce (remarque 5.4). Chaque équipe trace, à la main ou au tableur, le nuage de points d'une paire de variables différente, et décrit ce qu'elle voit. La mise en commun montre que certains nuages font apparaître des groupes, et d'autres non. On révèle alors les espèces, et l'on discute : pourquoi certaines vues sont-elles plus informatives ? Comment pourrait-on combiner deux mesures pour faire encore mieux ? En prolongement, les élèves peuvent essayer des combinaisons simples (somme ou différence de cotes z) et comparer les nuages obtenus.

c) Quelle proportion de la variance le plan représente-t-il ? (S'il n'en conserve que 40 %, les distances sur le graphique sont peu fiables.) Les variables ont-elles été réduites ? Qui a nommé les axes, et sur quelle base ? Le nom d'un axe est une *interprétation*, tirée des composantes du vecteur propre, et non un résultat du calcul. Le choix des vingt indicateurs, enfin, détermine entièrement le résultat. L'enseignant peut faire vivre ces questions en faisant construire aux élèves un « indice » combinant quelques indicateurs, puis en leur montrant qu'un autre choix de pondération modifie le classement : le résumé d'un ensemble de données n'est jamais neutre.

Chapitre 6 – Numération et stockage de l'information

Exercice 6.1.

a) 16
b) $128 + 32 + 8 + 2 = 170$
c) $8 + 4 + 2 + 1 = 15$
d) $255 - 4 = 251$

Exercice 6.2.

a) `00000000`
b) $21 = 16 + 4 + 1$: `00010101`
c) $144 = 128 + 16$: `10010000`
d) `11111011`

Exercice 6.3.

a) `A`
b) $100 = 6 \times 16 + 4$: `64`
c) $243 = 15 \times 16 + 3$: `F3`
d) $9999 = 2 \times 4096 + 7 \times 256 + 0 \times 16 + 15$: `270F`

Exercice 6.4.

a) $256 + 2 \times 16 + 3 = 291$
b) $256 + 10 \times 16 + 12 = 428$
c) $10 \times 4096 + 11 \times 256 + 11 \times 16 + 10 = 43\,962$
d) $9 \times 4096 + 12 \times 256 + 9 \times 16 + 10 = 40\,090$

Exercice 6.5.

a) Par blocs de 3 : 101 101 101 110, soit $(5556)_8$. Par blocs de 4 : 1011 0110 1110, soit $(\mathrm{B6E})_{16}$.
b) $(7\mathrm{EA})_{16} = (0111\,1110\,1010)_2$. En regroupant par 3, $(011\,111\,101\,010)_2 = (3752)_8$; c'est l'année 2026 (exemple 6.1).
c) Parce que 10 n'est pas une puissance de 2 : un chiffre décimal ne correspond pas à un nombre entier de bits. En revanche, $9 = 3^2$, et le raisonnement qui suit la proposition 6.2 s'adapte mot pour mot : chaque chiffre en base neuf correspond à un bloc de deux chiffres en base trois.

Exercice 6.6. Chaque chiffre hexadécimal correspond à 4 bits : il faut $16 \times 4 = 64$ bits, soit 8 octets. On peut représenter $16^{16} = 2^{64} \approx 1{,}8 \times 10^{19}$ valeurs.

Exercice 6.7.

a) Le nombre de chiffres est $\lfloor \log_b 10^{100} \rfloor + 1$. En base deux : $\lfloor 100 \log_2 10 \rfloor + 1 = \lfloor 332{,}19 \rfloor + 1 = 333$. En base seize : $\lfloor 100 \log_{16} 10 \rfloor + 1 = \lfloor 83{,}05 \rfloor + 1 = 84$.
b) Si n a k chiffres en base dix, alors $n < 10^k$, d'où $\log_2 n < k \log_2 10$, et le nombre de bits, $\lfloor \log_2 n \rfloor + 1$, est au plus $\lfloor k \log_2 10 \rfloor + 1$.

Exercice 6.8.

a) Si $n = \sum c_i 2^i$, tous les termes avec $i \geq 3$ sont divisibles par 8, donc $n \equiv (c_2 c_1 c_0)_2 \pmod 8$. Or $(c_2 c_1 c_0)_2$ est compris entre 0 et 7 : il est divisible par 8 si et seulement s'il est nul.
b) Comme $16 \equiv 1 \pmod{15}$, on a $16^i \equiv 1$ pour tout i, d'où $n = \sum c_i 16^i \equiv \sum c_i \pmod{15}$. C'est la généralisation de la preuve par neuf : en base b, un entier est congru à la somme de ses chiffres modulo $b - 1$.
c) La somme des chiffres est $10 + 11 + 11 + 10 = 42 \equiv 12 \pmod{15}$. Comme 3 et 5 divisent 15, on a aussi $n \equiv 42 \pmod 3$ et $n \equiv 42 \pmod 5$, soit $n \equiv 0 \pmod 3$ et $n \equiv 2 \pmod 5$: n est divisible par 3 mais pas par 5. En effet, $43\,962 = 3 \times 14\,654$.

Exercice 6.9.

a) Le développement est périodique, de période `0011`, à partir du deuxième bit ; tronquer après 23 bits revient à éliminer la partie qui suit le bit de rang 23. Or la suite des bits à partir du rang 24 est `0011 0011`…, qui est celle de 0,1 à partir du rang 4 ; décaler de 20 rangs divise par 2^{20}. L'erreur est donc $0{,}1 \times 2^{-20} \approx 9{,}54 \times 10^{-8}$ seconde.
b) En 100 heures, il y a $100 \times 3600 \times 10 = 3\,600\,000$ dixièmes de seconde, d'où une erreur de $3\,600\,000 \times 9{,}54 \times 10^{-8} \approx 0{,}34$ seconde.
c) $0{,}34 \times 1676 \approx 575$ mètres.

Exercice 6.10.

a) `#00FF00` est le vert pur ; `#000080` un bleu foncé (le *navy* ou bleu marine) ; `#FFC0CB` beaucoup de rouge et une quantité égale, assez grande, de vert et de bleu : un rose pâle ; `#C0C0C0` un gris clair (l'*argent*).

b) \$255 = \$ `FF`, \$165 = 10 ×16 + 5 = \$ `A5`, \$0 = \$ `00` : `#FFA500`, l'orange.

c) \$255 - 28 = 227 = \$ `E3`, \$255 - 63 = 192 = \$ `C0`, \$255 - 110 = 145 = \$ `91` : `#E3C091`, un beige orangé.

d) $16^3 = 4096$ couleurs.

e) Sans palette, $640 \times 480 \times 3 = 921\,600$ octets. Avec palette, $640 \times 480 \times 1 = 307\,200$ octets pour les pixels, plus $256 \times 3 = 768$ octets pour la palette, soit 307 968 octets. On économise 613 632 octets, soit environ deux tiers.

Exercice 6.11.

a) $(640)_8 = (110\,100\,000)_2$: le propriétaire peut lire et écrire (`rw-`), le groupe peut seulement lire (`r--`), les autres n'ont aucun droit (`---`).

b) `chmod 700`.

c) $2^9 = 512$, soit les nombres octaux de 000 à 777.

Exercice 6.12.

a) « ç » : $231 = (000\,1110\,0111)_2$, réparti en `00011` et `100111`, d'où `11000011 10100111`, soit `C3 A7`. « « » : $171 = (000\,1010\,1011)_2$, d'où `11000010 10101011`, soit `C2 AB`. « œ » : $339 = (001\,0101\,0011)_2$, réparti en `00101` et `010011`, d'où `11000101 10010011`, soit `C5 93`.

b) `43`, `61` et `66` sont des caractères ASCII : « C », « a », « f ». `C3 A9` est « é » (exemple 6.3). On lit « Café ».

c) « Ç » occupe 2 octets, les cinq autres caractères (« a », l'espace, « v », « a », « ? ») un seul : 7 octets en UTF-8, contre 6 en Windows-1252.

d) « ç » est `C3 A7`, lu comme « Ã » et « § » : on affiche « garÃ§on ».

e) $3 + 6 + 6 + 6 = 21$ bits, soit $2^{21} = 2\,097\,152$ valeurs. Or `10FFFF` $= 1\,114\,111 < 2^{21}$: il s'écrit sur 21 bits.

Exercice 6.13.

a) $\frac{2\times10^{12}}{2^{30}} \approx 1862{,}6$: Windows affiche environ 1862 « Go », ou 1,82 « To ».

b) 4 Go $= 32 \times 10^9$ bits, et $\frac{32\times10^9}{50\times10^6} = 640$ secondes, soit un peu plus de 10 minutes.

c) Les photos occupent 45 000 Mo, soit 45 Go : oui, avec une bonne marge.

d) L'écart vaut $\left(\frac{1024}{1000}\right)^k - 1$. On cherche $1{,}024^k > 1{,}2$, soit $k > \frac{\ln 1{,}2}{\ln 1{,}024} \approx 7{,}7$: à partir de $k = 8$, les yottaoctets (10^{24}) et les yobioctets (2^{80}), l'écart dépasse 20 % (20,9 %).

Exercice 6.14. L'image compte $1024 \times 1280 = 1\,310\,720$ pixels.

a) Un bit par pixel : 1 310 720 bits, soit 163 840 octets.

b) Comme $256 = 2^8$, un octet par pixel : 1 310 720 octets.

c) Trois octets par pixel : 3 932 160 octets.

d) $3\,932\,160 \div 4 = 983\,040$ octets.

Exercice 6.15.

a) 4 blancs, 8 noirs, 11 blancs, 2 noirs : « 4, 8, 11, 2 ».

b) La ligne commence par 0 blanc : `1110000011`.

c) Une ligne qui alterne : `0101010101010101`, codée « 1, 1, 1, … , 1 », par 16 nombres, ce qui est plus long que la ligne elle-même.

Exercice 6.16.

a) Une image matricielle est décrite pixel par pixel ; une image vectorielle, par des objets géométriques et leurs caractéristiques. L'image vectorielle est recalculée à chaque affichage et peut être agrandie sans perte de qualité, contrairement à l'image matricielle. Exemples : photos et images numérisées (matricielles) ; logos, plans, figures géométriques, polices de caractères (vectorielles).

b) Le format BMP stocke les données sans compression : chaque pixel occupe le même nombre de bits. Le format JPEG utilise une compression avec perte, qui décrit chaque bloc de 8×8 pixels par ses motifs grossiers et élimine une partie des détails fins, peu visibles à l'œil.

c) En SVG : la figure est faite d'objets géométriques, et une image vectorielle s'agrandit sans perte. Une image PNG assez fine pour une affiche serait très volumineuse, et ses contours paraîtraient crénelés à une définition plus faible.

Exercice 6.17.

```
def ecrire(n, b):
    if n == 0:
        return [0]
    chiffres = []
    while n > 0:
        n, r = divmod(n, b)      # division euclidienne
        chiffres.append(r)
    return chiffres[::-1]        # restes lus du dernier au premier

def lire(chiffres, b):
    n = 0
    for c in chiffres:
        n = n * b + c            # schéma de Horner
    return n

for b in (2, 8, 16):
    print(b, ecrire(2026, b), lire(ecrire(2026, b), b))
print(bin(2026), oct(2026), hex(2026))
```

On obtient `[1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0]`, `[3, 7, 5, 2]` et `[7, 14, 10]`, c'est-à-dire `0b11111101010`, `0o3752` et `0x7ea`.

Exercice 6.18. Les opérations `p >> k` (décalage de k bits vers la droite, soit le quotient par 2^k) et `p & 0b111111` (les 6 derniers bits, soit le reste modulo 64) extraient les blocs de bits.

```
def utf8(p):
    if p < 0x80:
        return [p]
    if p < 0x800:
        return [0b11000000 | (p >> 6),
                0b10000000 | (p & 0b111111)]
    if p < 0x10000:
        return [0b11100000 | (p >> 12),
                0b10000000 | ((p >> 6) & 0b111111),
```

```
                0b10000000 | (p & 0b111111)]
    return [0b11110000 | (p >> 18),
            0b10000000 | ((p >> 12) & 0b111111),
            0b10000000 | ((p >> 6) & 0b111111),
            0b10000000 | (p & 0b111111)]

for c in ["A", "é", "€", chr(0x1F600)]:
    print(c, bytes(utf8(ord(c))).hex(" "), c.encode("utf-8").hex(" "))
```

Exercice 6.19.

a)

```
with open("degrade.ppm", "w") as f:
    f.write("P3\n256 256\n255\n")
    for y in range(256):
        for x in range(256):
            f.write(f"{x} {y} 128\n")
```

b) Les données brutes occupent 196 608 octets. Le fichier texte en occupe 730 127, près de quatre fois plus : chaque composante y est écrite en caractères ASCII, avec jusqu'à 3 chiffres et un séparateur, au lieu d'un seul octet. C'est le prix de la lisibilité. Le format binaire `P6` stocke au contraire chaque composante sur un octet.

Exercice 6.20.

Plusieurs réponses sont possibles ; en voici une.

a) La *notation exponentielle et les puissances* : elles sont au cœur de la numération positionnelle et des unités de stockage (2^{10}, 10^3). Le *dénombrement* (principe multiplicatif) : il donne le nombre de valeurs représentables avec n bits, ou le nombre de couleurs RVB. Les *fonctions exponentielles et logarithmiques* (cinquième secondaire, TS et SN) : elles donnent le nombre de chiffres d'un entier dans une base et décrivent la croissance des capacités de stockage. On pourrait ajouter les *fractions et leur développement décimal*, à propos de l'arrondi de 0,1.

b) « Ton idée est bonne, mais incomplète. Dans un circuit, on ne regarde pas si le courant passe, mais si la tension est basse ou haute ; on choisit deux niveaux très éloignés pour ne pas les confondre, même s'il y a du bruit. Sur un disque, ce n'est pas du courant, mais de petits aimants orientés dans un sens ou dans l'autre ; sur une clé USB, des électrons emprisonnés. Et les 0 et les 1 ne veulent rien dire seuls : il faut une convention, comme un code secret, pour savoir si `01000001` est le nombre 65, la lettre « A » ou un morceau de couleur. »

c) On projette des carrés colorés accompagnés de leur code (`#FF0000`, `#00FF00`, `#FFFF00`, `#808080`…), et les élèves formulent des hypothèses sur la signification des trois paires. Ils découvrent que chaque paire va de `00` à `FF` et doivent déterminer combien de valeurs cela représente. La question « que vient-il après `09` ? après `0F` ? » amène à expliciter la règle de position en base seize, par analogie avec l'odomètre d'une voiture. On conclut par un défi : produire une couleur donnée à l'aide d'un sélecteur de couleur, puis prédire le code d'une couleur sans l'afficher.

Chapitre 7 – La cryptographie

Exercice 7.1.

a) `JXLWDUH`.

b) On recule de trois rangs : `auto`.

c) La lettre c (2) est devenue H (7) : la clé est $k = 5$. Le mot `chat` devient `HMFY`.

d) Appliquer j fois le chiffre de César revient à décaler de $3j$. On retrouve le texte de départ lorsque $3j \equiv 0 \pmod{26}$, soit, puisque $\operatorname{pgcd}(3, 26) = 1$, lorsque $26 \mid j$: il faut l'appliquer 26 fois. Pour une clé k, il faut le plus petit $j \geq 1$ tel que $26 \mid jk$, c'est-à-dire $j = 26/\operatorname{pgcd}(k, 26)$. Par exemple, $j = 2$ pour $k = 13$ (ROT13) et $j = 13$ pour $k = 2$.

Exercice 7.2.

a) `KDGNHJFVCYSMT`.

b) Si $\ell = 1$, on a $i \bmod \ell = 0$ pour tout i, et chaque lettre est décalée de k_0.

c) Soient $(k_0, \dots, k_{\ell_1-1})$ et $(k'_0, \dots, k'_{\ell_2-1})$ les deux clés. La lettre en position i est décalée de $s_i = k_{i \bmod \ell_1} + k'_{i \bmod \ell_2}$. Posons $L = \operatorname{ppcm}(\ell_1, \ell_2)$. Comme ℓ_1 et ℓ_2 divisent L, on a $(i + L) \bmod \ell_1 = i \bmod \ell_1$ et de même pour ℓ_2, d'où $s_{i+L} = s_i$: le résultat est un chiffre de Vigenère de clé $(s_0, \dots, s_{L-1})$. Avec $\ell_1 = 5$ et $\ell_2 = 6$, on obtient une clé de longueur 30, plus longue que chacune des deux, ce qui complique l'attaque de la section 7.5. Mais si $\ell_1 = \ell_2$, on ne gagne rien du tout, et la méthode de Friedman s'applique de toute façon, avec un texte plus long.

Exercice 7.3.

a) `UCSAVXIWPC`.

b) L'inverse de 5 modulo 26 est 21, car $5 \times 21 = 105 = 4 \times 26 + 1$. La clé de déchiffrement est $(21, -21 \times 8) = (21, -168) \equiv (21, 14)$. On obtient `pythagore`.

c) Chiffrer deux fois donne $\alpha(\alpha x + \beta) + \beta = \alpha^2 x + (\alpha + 1)\beta$. Il faut que $\alpha^2 x + (\alpha+1)\beta \equiv x$ pour tout x. Avec $x = 0$, on obtient $(\alpha+1)\beta \equiv 0$, puis, avec $x = 1$, $\alpha^2 \equiv 1 \pmod{26}$. En essayant les douze valeurs admissibles de α, on trouve $\alpha = 1$ et $\alpha = 25$. Si $\alpha = 1$, il faut $2\beta \equiv 0$, soit $\beta = 0$ ou $\beta = 13$. Si $\alpha = 25$, $(\alpha + 1)\beta = 26\beta \equiv 0$ pour tout β. Il y a donc 28 involutions : $(1, 0)$, $(1, 13)$ et les 26 clés $(25, \beta)$, qui correspondent à $x \mapsto \beta - x$.

Exercice 7.4. On a $e = 4 \mapsto M = 12$ et $u = 20 \mapsto U = 20$, d'où $4\alpha + \beta \equiv 12$ et $20\alpha + \beta \equiv 20 \pmod{26}$. En soustrayant, $16\alpha \equiv 8 \pmod{26}$. Comme $\operatorname{pgcd}(16, 26) = 2$ divise 8, on divise par 2 : $8\alpha \equiv 4 \pmod{13}$. L'inverse de 8 modulo 13 est 5 ($8 \times 5 = 40 = 3 \times 13 + 1$), d'où $\alpha \equiv 20 \equiv 7 \pmod{13}$, soit $\alpha = 7$ ou $\alpha = 20$ modulo 26. Seule la valeur impaire est admissible : $\alpha = 7$, puis $\beta \equiv 12 - 28 = -16 \equiv 10$. La clé de déchiffrement est $(15, -150) \equiv (15, 6)$, et le message est `eureka`.

Exercice 7.5.

a) Seul `camille` contient une lettre doublée, et seul `DHJRYYZ` aussi : `camille` $\mapsto$ `DHJRYYZ`. On en déduit que $c \mapsto D$, $a \mapsto H$, $m \mapsto J$, $i \mapsto R$, $l \mapsto Y$ et $e \mapsto Z$. Le seul prénom de cinq lettres est `felix` $\mapsto$ `KZYRA`. Parmi les prénoms de six lettres, `samuel` commence par une lettre dont l'image ($a \mapsto H$ en deuxième position) permet de le reconnaître : `samuel` $\mapsto$ `XHJTZY` et `noemie` $\mapsto$ `UFZJRZ`. Enfin, `olivier` $\mapsto$ `FYRERZM` (le i apparaît deux fois, comme le R) et `rosalie` $\mapsto$ `MFXHYRZ`.

b) Comme $a = 0 \mapsto H = 7$, on a directement $\beta = 7$. Puis $c = 2 \mapsto D = 3$ donne $2\alpha + 7 \equiv 3$, soit $2\alpha \equiv 22 \pmod{26}$, d'où $\alpha \equiv 11 \pmod{13}$: $\alpha = 11$ ou $\alpha = 24$, et seule la valeur impaire convient. La clé est $(11, 7)$. L'inverse de 11 modulo 26 est 19, et la clé de déchiffrement est $(19, -133) \equiv (19, 23)$.

Exercice 7.6.

a) Par le principe multiplicatif, $26! \approx 4{,}03 \times 10^{26}$.

b) Oui : par la proposition 7.1, la fonction $x \mapsto \alpha x + \beta$ est une bijection de $\mathbb{Z}_{26}$ lorsque $\operatorname{pgcd}(\alpha, 26) = 1$.

c) Non : il n'y a que 312 chiffres affines, beaucoup moins que 26! permutations. Pour la permutation donnée, $a = 0 \mapsto g = 6$ forcerait $\beta = 6$, puis $b = 1 \mapsto f = 5$ forcerait $\alpha + 6 \equiv 5$, soit $\alpha = 25$. Le chiffre affine $x \mapsto 25x + 6$ envoie bien c, d, e, f, g sur e, d, c, b, a, mais il envoie $h = 7$ sur $25 \times 7 + 6 = 181 \equiv 25$, soit z, alors que la permutation fixe h. Ce n'est donc pas un chiffre affine.

d) La lettre x est inchangée si $\alpha x + \beta \equiv x$, soit $(\alpha - 1)x \equiv -\beta \pmod{26}$. Comme α est impair, $\alpha - 1$ est pair, et $\operatorname{pgcd}(\alpha - 1, 26)$ vaut 2, sauf si $\alpha = 1$, auquel cas il vaut 26 (il ne peut pas valoir 13, qui est impair). Par la proposition E.20 :
— si $\alpha = 1$ (un décalage), aucune lettre n'est inchangée, sauf si $\beta = 0$, auquel cas toutes le sont ;
— si $\alpha \neq 1$ et β est impair, aucune lettre n'est inchangée ;
— si $\alpha \neq 1$ et $\beta = 2j$ est pair, la congruence devient $\frac{\alpha-1}{2}\, x \equiv -j \pmod{13}$, qui a une seule solution x_0 modulo 13 : exactement deux lettres sont inchangées, x_0 et $x_0 + 13$.
Par exemple, pour la clé $(7, 4)$, on résout $3x \equiv -2 \equiv 11 \pmod{13}$; l'inverse de 3 est 9, d'où $x \equiv 99 \equiv 8$: les lettres i (8) et v (21) sont inchangées. Contrairement à l'Enigma (proposition 7.3), le chiffre affine laisse donc souvent des lettres en place.

Exercice 7.7.

a) $\det K = 8 + 1 = 9$, et $K^{-1} = \frac{1}{9}\begin{pmatrix} 2 & -1 \\ 1 & 4 \end{pmatrix}$ dans $\mathbb{R}$. Comme $\operatorname{pgcd}(9, 26) = 1$, K est inversible modulo 26 (théorème 7.2). L'inverse de 9 modulo 26 est 3 ($27 \equiv 1$), d'où

$$K^{-1} \equiv 3\begin{pmatrix} 2 & -1 \\ 1 & 4 \end{pmatrix} = \begin{pmatrix} 6 & -3 \\ 3 & 12 \end{pmatrix} \equiv \begin{pmatrix} 6 & 23 \\ 3 & 12 \end{pmatrix} \pmod{26}.$$

b) Le premier bloc, WO $= (22, 14)$, donne $(6 \times 22 + 23 \times 14,\ 3 \times 22 + 12 \times 14) = (454, 234) \equiv (12, 0)$, soit ma. On obtient mathematiquesx : le x final complète le dernier bloc.

c) $X = \begin{pmatrix} 19 & 0 \\ 7 & 19 \end{pmatrix}$ et $Y = \begin{pmatrix} 5 & 19 \\ 21 & 12 \end{pmatrix}$. On a $\det X = 361 \equiv 23 \pmod{26}$, d'inverse 17 ($23 \times 17 = 391 = 15 \times 26 + 1$), d'où $X^{-1} \equiv 17\begin{pmatrix} 19 & 0 \\ -7 & 19 \end{pmatrix} \equiv \begin{pmatrix} 11 & 0 \\ 11 & 11 \end{pmatrix}$ et

$$YX^{-1} = \begin{pmatrix} 264 & 209 \\ 363 & 132 \end{pmatrix} \equiv \begin{pmatrix} 4 & 1 \\ 25 & 2 \end{pmatrix} \pmod{26},$$

qui est bien K, puisque $-1 \equiv 25$.

d) Avec ma $= (12, 0)$ et th $= (19, 7)$, la matrice $X = \begin{pmatrix} 12 & 19 \\ 0 & 7 \end{pmatrix}$ a pour déterminant $84 \equiv 6 \pmod{26}$, qui n'est pas premier avec 26 : X n'est pas inversible modulo

26, et l'équation $Y \equiv KX$ a plusieurs solutions. Cherchons une ligne (r_1, r_2), non nulle, telle que $(r_1, r_2)\,X \equiv (0,0)$: il faut $12r_1 \equiv 0$, d'où $r_1 = 13$ (ou 0), puis $19 \times 13 + 7r_2 \equiv 0$, soit $7r_2 \equiv 13$, d'où $r_2 = 13$. La matrice $K' = K + \begin{pmatrix} 13 & 13 \\ 0 & 0 \end{pmatrix} \equiv \begin{pmatrix} 17 & 14 \\ 25 & 2 \end{pmatrix}$ envoie donc elle aussi `ma` sur `WO` et `th` sur `FV`. Ces deux blocs ne permettent pas de distinguer K de K'.

Exercice 7.8.

a) Comme p est premier, tout élément non nul de $\mathbb{Z}_p$ est inversible, et une matrice est inversible modulo p si et seulement si son déterminant n'est pas nul modulo p. Si la première colonne (a, c) est nulle, ou si la seconde vaut $\lambda(a, c)$, le déterminant est $ad - bc = a\lambda c - \lambda ac = 0$. Réciproquement, si $ad - bc \equiv 0$ et $(a, c) \neq (0, 0)$, disons $a \neq 0$, alors $(b, d) = \lambda(a, c)$ avec $\lambda = ba^{-1}$, puisque $\lambda c = ba^{-1}c = a^{-1}(ad) = d$ (le cas $c \neq 0$ est semblable). Il y a $p^2 - 1$ choix pour la première colonne, et, une fois celle-ci choisie, $p^2 - p$ choix pour la seconde (tous les vecteurs sauf les p multiples de la première).

b) Par le théorème des restes chinois, se donner un coefficient modulo 26, c'est se donner son reste modulo 2 et son reste modulo 13 ; une matrice modulo 26 correspond donc à un couple formé d'une matrice modulo 2 et d'une matrice modulo 13. Par exemple,

$$B = \begin{pmatrix} 15 & 4 \\ 20 & 9 \end{pmatrix} \quad \longmapsto \quad \left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \bmod 2,\ \begin{pmatrix} 2 & 4 \\ 7 & 9 \end{pmatrix} \bmod 13 \right),$$

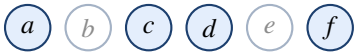
puisque $15 = 7 \times 2 + 1 = 13 + 2$, $4 = 2 \times 2 + 0$, $20 = 10 \times 2 + 0 = 13 + 7$ et $9 = 4 \times 2 + 1$. Inversement, le couple de droite détermine B : le coefficient en haut à gauche est l'unique entier de $\{0, \dots, 25\}$ qui est impair et congru à 2 modulo 13, c'est-à-dire 15 (et non 2), et de même pour les trois autres. Les déterminants se correspondent aussi : $\det B = 135 - 80 = 55$, et l'on a bien $55 \equiv 1 = \det \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \pmod 2$ et $55 \equiv 3 \equiv 18 - 28 = \det \begin{pmatrix} 2 & 4 \\ 7 & 9 \end{pmatrix} \pmod{13}$. Les deux matrices de droite sont inversibles, et B l'est aussi, puisque $\operatorname{pgcd}(55, 26) = 1$. En revanche, $C = \begin{pmatrix} 4 & 1 \\ 6 & 3 \end{pmatrix}$ donne $\begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$ modulo 2, de déterminant nul : C n'est pas inversible modulo 26 ($\det C = 6$ est pair), même si elle l'est modulo 13.

En général, une matrice A est inversible modulo 26 si et seulement si $\operatorname{pgcd}(\det A, 26) = 1$, c'est-à-dire si $\det A$ n'est divisible ni par 2 ni par 13 : si et seulement si elle est inversible modulo 2 et modulo 13. Par la partie a), il y a $(4-1)(4-2) = 6$ matrices inversibles modulo 2 et $(169-1)(169-13) = 26\,208$ modulo 13, d'où $6 \times 26\,208 = 157\,248$ clés, soit environ 34 % des $26^4 = 456\,976$ matrices.

Exercice 7.9.

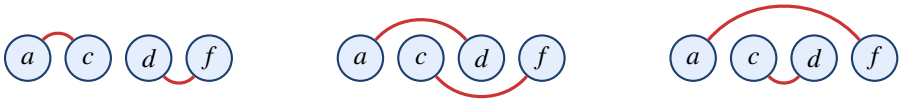
a) Appelons les six lettres a, b, c, d, e, f. Chaque câble relie deux lettres : avec deux câbles, exactement 4 des 6 lettres sont branchées, et les 2 autres restent seules (elles sont chiffrées sans être échangées). On compte les branchements en deux étapes, illustrées ci-dessous.

Étape 1 : quelles lettres reçoivent un câble ?



4 lettres reliées parmi 6 : $\binom{6}{4} = 15$ choix
(ici a, c, d, f ; b et e restent seules)

Étape 2 : comment apparier a, c, d, f ?



3 appariements possibles, d'où $15 \times 3 = 45$ branchements.

Étape 1 : quelles lettres sont branchées ? Il faut choisir 4 lettres parmi 6, sans tenir compte de l'ordre : il y a $\binom{6}{4} = 15$ façons de le faire. (On peut aussi choisir les 2 lettres qui restent seules : $\binom{6}{2} = 15$, le même nombre.) Le coefficient binomial apparaît parce qu'à cette étape, seul compte l'*ensemble* des lettres branchées, pas encore la façon de les relier.

Étape 2 : comment relier ces 4 lettres ? Une fois les lettres choisies, disons a, c, d, f, le partenaire de a détermine tout : si a est reliée à c, le second câble relie forcément d et f. Il y a donc 3 façons d'apparier les 4 lettres : $\{a, c\}\{d, f\}$, $\{a, d\}\{c, f\}$ et $\{a, f\}\{c, d\}$.

Par le principe multiplicatif, il y a $15 \times 3 = 45$ branchements. La formule de b) donne bien $6!/(2! \cdot 2! \cdot 2^2) = 720/16 = 45$.

b) On choisit les $2k$ lettres reliées, de $\binom{26}{2k}$ façons. On les range ensuite dans un certain ordre ($(2k)!$ façons), et l'on relie la première à la deuxième, la troisième à la quatrième, et ainsi de suite. Chaque branchement est alors obtenu $k!\,2^k$ fois : on peut permuter les k paires, et échanger les deux lettres de chaque paire. Ainsi $N_k = \binom{26}{2k}\frac{(2k)!}{k!\,2^k} = \frac{26!}{(26-2k)!\,k!\,2^k}$.

c) $N_9 \approx 5{,}4 \times 10^{13}$, $N_{10} \approx 1{,}51 \times 10^{14}$, $N_{11} \approx 2{,}06 \times 10^{14}$, $N_{12} \approx 1{,}03 \times 10^{14}$ et $N_{13} \approx 7{,}9 \times 10^{12}$. Le maximum est atteint pour $k = 11$. Avec 13 câbles, toutes les lettres sont reliées, et il reste moins de choix : il n'y a plus qu'à décider *comment* apparier les lettres, et non *lesquelles*. Les Allemands utilisaient 10 câbles, presque l'optimum.

Exercice 7.10.

a) $n = 119$ et $\varphi(n) = 6 \times 16 = 96 = 2^5 \times 3$, premier avec 5. On a $96 = 19 \times 5 + 1$, d'où $1 = 96 - 19 \times 5$ et $5 \times (-19) \equiv 1 \pmod{96}$: $d = -19 + 96 = 77$. On vérifie : $5 \times 77 = 385 = 4 \times 96 + 1$.

b) $10^2 = 100$, $10^4 \equiv 100^2 = 10\,000 = 84 \times 119 + 4 \equiv 4$, d'où $10^5 \equiv 40 \pmod{119}$: $c = 40$.

c) $77 = 64 + 8 + 4 + 1$. On calcule les carrés successifs modulo 119 : $40^2 = 1600 \equiv 53$, $40^4 \equiv 53^2 = 2809 \equiv 72$, $40^8 \equiv 72^2 = 5184 \equiv 67$, $40^{16} \equiv 67^2 = 4489 \equiv 86$, $40^{32} \equiv 86^2 = 7396 \equiv 18$, $40^{64} \equiv 18^2 = 324 \equiv 86$. Alors $40^{77} = 40^{64} \cdot 40^8 \cdot 40^4 \cdot 40 \equiv 86 \times 67 \times 72 \times 40$. On a $86 \times 67 = 5762 \equiv 50$, $50 \times 72 = 3600 \equiv 30$ et $30 \times 40 = 1200 \equiv 10 \pmod{119}$. On retrouve $m = 10$.

d) Elle factorise 119 : en essayant les nombres premiers jusqu'à $\sqrt{119} < 11$, elle trouve $119 = 7 \times 17$. Elle calcule alors $\varphi(n) = 96$ et d comme en a).

Exercice 7.11.

a) On a $b \equiv c^e \equiv (m^e)^e = m^{e^2} \pmod{n}$. Comme $e^2 = 1 + k\varphi(n)$ pour un certain entier $k \geq 0$, le théorème d'Euler donne $b \equiv m \cdot \left(m^{\varphi(n)}\right)^k \equiv m \pmod{n}$: le message chiffré deux fois est le message clair. Autrement dit, e est son propre inverse modulo $\varphi(n)$: la clé de chiffrement est aussi la clé de déchiffrement.

b) Ici $\varphi(n) = 2 \times 10 = 20$. Les exposants admissibles, premiers avec 20 et compris entre 2 et 19, sont $3, 7, 9, 11, 13, 17, 19$. Parmi eux, $9^2 = 81$, $11^2 = 121$ et $19^2 = 361$ sont congrus à 1 modulo 20, alors que $3^2 = 9$, $7^2 = 49$, $13^2 = 169$ et $17^2 = 289$ ne le sont pas. Il fallait éviter 9, 11 et 19 (ainsi que $e = 1$, qui ne chiffre rien).

Exercice 7.12.

a) Comme $\text{pgcd}(e_M, e_G) = 1$, le théorème de Bachet–Bézout (théorème E.9) fournit des entiers x et y tels que $e_M x + e_G y = 1$, que Dimitri calcule par l'algorithme d'Euclide étendu, puisque e_M et e_G sont publics. L'un des deux est négatif, disons $x < 0$. Comme m est premier avec n, c_M l'est aussi, et possède un inverse c_M^{-1} modulo n. Dimitri calcule

$$\left(c_M^{-1}\right)^{-x} c_G^y \equiv m^{e_M x}\, m^{e_G y} = m^{e_M x + e_G y} = m \pmod{n}.$$

Le module ne doit donc jamais être partagé entre deux personnes.

b) $10 = 1\times 7 + 3$ et $7 = 2\times 3 + 1$, d'où $1 = 7 - 2\times 3 = 7 - 2\,(10 - 7) = 3\times 7 - 2\times 10$. On prend $x = -2$ et $y = 3$, et Dimitri calcule $m \equiv \left(c_M^{-1}\right)^2 c_G^3 \pmod{n}$.

Exercice 7.13.

a) Écrivons $p = 3a + 2$ et $q = 3b + 2$. Alors $2(p-1)(q-1) + 1 = 2(3a+1)(3b+1) + 1 = 18ab + 6a + 6b + 3$, et $d = 6ab + 2a + 2b + 1$ est un entier.

b) Si $y \equiv 0 \pmod{n}$, alors $n \mid x^3$, et le nombre premier p diviserait x^3, donc x (lemme d'Euclide, corollaire E.12) ; c'est impossible, puisque $\text{pgcd}(x, n) = 1$. De même, $z \equiv x^{3d} \pmod{n}$, et $n \mid x^{3d}$ est impossible.

c) On a $3d = 2(p-1)(q-1) + 1 = 2\varphi(n) + 1$, donc $z \equiv x^{3d} = x \cdot \left(x^{\varphi(n)}\right)^2 \equiv x \pmod{n}$ par le théorème d'Euler. C'est le chiffrement RSA avec l'exposant $e = 3$: la condition $p \equiv q \equiv 2 \pmod{3}$ garantit que 3 ne divise ni $p - 1$ ni $q - 1$, donc que $\text{pgcd}(3, \varphi(n)) = 1$, et d est précisément l'inverse de 3 modulo $\varphi(n)$, puisque $3d = 1 + 2\varphi(n)$.

Exercice 7.14.

a) $A = 5^6 = 15\,625 \equiv 8 \pmod{23}$ et $B = 5^{15} \equiv 19 \pmod{23}$ (par exemple, $5^2 \equiv 2$, $5^4 \equiv 4$, $5^8 \equiv 16$, et $5^{15} = 5^8 \cdot 5^4 \cdot 5^2 \cdot 5 \equiv 16 \times 4 \times 2 \times 5 = 640 \equiv 19$). La clé commune est $19^6 \equiv 8^{15} \equiv 2 \pmod{23}$.

b) Les puissances de 5 modulo 23 sont $5, 2, 10, 4, 20, 8, \ldots$: $5^6 \equiv 8$, d'où $a = 6$. Avec p de plusieurs centaines de chiffres, cette liste serait beaucoup trop longue pour être calculée.

c) Si $g = 1$, toutes les puissances valent 1, et la clé commune est toujours 1. Si $g = 22 \equiv -1$, les puissances valent 1 ou -1 : la clé commune ne peut prendre que deux valeurs. On choisit g de sorte que ses puissances prennent un très grand nombre de valeurs. Ici, 5 est un bon choix : ses puissances parcourent les 22 valeurs de 1 à 22.

Exercice 7.15. On utilise les fonctions des programmes 7.1 et 7.2. Comme le déchiffrement est lui-même un chiffre affine (proposition 7.1), on parcourt les clés de déchiffrement directement.

```
def score(texte):
    return sum(FREQ[x] for x in vers_nombres(texte))

def casser_affine(chiffre):
    alphas = [a for a in range(1, 26) if a % 2 == 1 and a != 13]
    cles = [(a, b) for a in alphas for b in range(26)]
    a, b = max(cles, key=lambda cle: score(affine(chiffre, *cle)))
    return (a, b), affine(chiffre, a, b)
```

Pour le texte de l'exemple 7.6, la fonction renvoie la clé de déchiffrement $(3, 10)$ et le texte `LACRYPTOGRAPHIEEST`…. Pour `PCRZCQJ`, elle renvoie `RETBESL`, et pour `MUZMCK`, `EAREWS` : sur des textes de quelques lettres, les fréquences ne veulent rien dire, et le texte « le plus français » au sens du score n'est pas le bon. Il faudrait alors un dictionnaire, ou un texte clair connu. Les méthodes statistiques demandent des textes assez longs.

Exercice 7.16.

```
import matplotlib.pyplot as plt

chiffre = "..."                          # le texte reçu
d = range(1, 21)
plt.bar(d, [coincidences(chiffre, k) for k in d])
plt.xlabel("décalage"); plt.ylabel("coïncidences")
plt.show()

longueur = 6                             # lue sur le diagramme
cle = trouver_cle(chiffre, longueur)
print(cle, vigenere(chiffre, cle, -1))
```

Les pics apparaissent aux multiples de la longueur de la clé. Avec le texte de l'exemple 7.6 (360 lettres) et la clé `secondaire`, chaque colonne ne compte que 36 lettres, et la méthode fonctionne encore. En général, elle commence à se tromper sur quelques lettres de la clé lorsque les colonnes comptent moins d'une vingtaine de lettres ; on corrige alors à la main, en regardant le texte déchiffré.

Exercice 7.17.

```
ROTOR = "ekmflgdqvzntowyhxuspaibrcj"
REFLECTEUR = "yruhqsldpxngokmiebfzcwvjat"
R = vers_nombres(ROTOR)
R_inv = [R.index(i) for i in range(26)]        # la permutation ré
    ciproque
U = vers_nombres(REFLECTEUR)

print(all(U[U[i]] == i and U[i] != i for i in range(26)))   # a) True

def mini_enigma(texte, depart=0):
    y = []
    for i, x in enumerate(vers_nombres(texte)):
        k = (depart + i + 1) % 26              # le rotor avance d'un
    cran
        z = (R[(x + k) % 26] - k) % 26         # aller dans le rotor
        z = U[z]                               # réflecteur
        z = (R_inv[(z + k) % 26] - k) % 26     # retour dans le rotor
        y.append(z)
```

```
    return vers_lettres(y)

print(mini_enigma("a" * 26))          # NRNLINTKBNJTOICVXPTJIQMODH
c = mini_enigma("attaque au bas du pont", 5)
print(c, mini_enigma(c, 5))           # NAQBSEGOJLVOVBIJKL
    ATTAQUEAUBASDUPONT
```

c) La même lettre a est chiffrée de façons différentes à chaque frappe : c'est un chiffre polyalphabétique. On ne voit jamais la lettre A dans le chiffré. Avec un seul rotor, la substitution se répète toutes les 26 lettres, d'où l'intérêt des trois rotors de la vraie machine.

d) Chiffrer deux fois, à partir de la même position, redonne le texte clair, et aucune lettre du chiffré ne coïncide avec la lettre claire correspondante : `any(a == b for a, b in zip(clair, chiffre))` renvoie `False`.

Exercice 7.18.

```
n, e, c = 1500011500021, 65537, 213152470940

def plus_petit_diviseur(n):
    d = 2
    while d * d <= n:
        if n % d == 0:
            return d
        d += 1
    return n                          # n est premier

p = plus_petit_diviseur(n)
q = n // p                            # p = 1000003, q = 1500007
d = pow(e, -1, (p - 1) * (q - 1))     # d = 462703235489
m = pow(c, d, n)                      # m = 42017041000
x = []
while m > 0:                          # on lit les lettres deux
    chiffres à la fois
    m, r = divmod(m, 100)
    x.append(r)
print(vers_lettres(x[::-1]))          # EUREKA
```

a) $n = 1\,000\,003 \times 1\,500\,007$.

b) $d = 462\,703\,235\,489$ et le message est `eureka`. (Si le message commençait par $a = 00$, ce zéro initial serait perdu ; on convient en général de commencer par une lettre autre que a, ou d'indiquer la longueur du message.)

c) La fonction a essayé environ un million de diviseurs, ce qui prend moins d'une seconde. Pour un nombre de 600 chiffres, il faudrait essayer de l'ordre de 10^{300} diviseurs, soit 10^{291} secondes, alors que l'âge de l'Univers est d'environ 4×10^{17} secondes. Les meilleurs algorithmes connus sont beaucoup plus rapides que les divisions successives, mais restent eux aussi hors d'atteinte pour de tels nombres.

Exercice 7.19. Plusieurs réponses sont possibles ; en voici une.

a) La *fonction affine et sa réciproque* : le chiffre affine et son déchiffrement (section 7.3). La *résolution de systèmes d'équations* : l'attaque à texte clair connu contre le chiffre affine et le déchiffrement de Hill. La *statistique descriptive* (tableaux de distribution, fréquences relatives, diagrammes à bandes) : l'analyse

des fréquences (section 7.4). Le *dénombrement* et le principe multiplicatif : le nombre de clés des substitutions, du chiffre affine et de l'Enigma. On pourrait aussi citer les nombres premiers et la factorisation (RSA), les lois des exposants (exponentiation rapide) et la notation scientifique (ordres de grandeur des nombres de clés).

b) *Intention* : faire découvrir qu'un grand nombre de clés ne suffit pas à protéger un message, et réinvestir la statistique descriptive. *Première période* : les élèves fabriquent une roue de César, chiffrent des messages et cassent ceux de leurs camarades en essayant les 26 clés ; on discute du nombre de clés, puis on passe aux substitutions quelconques (26! clés, calculées à la calculatrice). *Seconde période* : chaque équipe reçoit un texte d'environ 300 lettres chiffré par substitution ; elle compte les lettres, construit un diagramme à bandes, le compare à celui du français et reconstitue le texte. Les élèves remettent le texte déchiffré et un court paragraphe expliquant leur démarche.

c) « Une méthode finit toujours par être connue : une machine est volée, un document est perdu, quelqu'un parle. Pendant la Seconde Guerre mondiale, les Allemands pensaient l'Enigma inviolable, mais les Polonais en ont reconstitué le fonctionnement, et des documents ont été saisis sur des navires. Dès 1883, Auguste Kerckhoffs recommandait de supposer que l'ennemi connaît la méthode, et de ne compter que sur le secret de la clé, beaucoup plus facile à changer. C'est d'ailleurs ce que l'on fait aujourd'hui : les méthodes qui protègent Internet sont publiques, et c'est parce que des milliers de chercheurs ont essayé en vain de les casser qu'on leur fait confiance. »

Annexe D – Rappels d'algèbre linéaire

Exercice D.1.

a) $2\mathbf{u} - \mathbf{v} = (2-3,\, -4-0,\, 4+4) = (-1,-4,8)$; $\|\mathbf{u}\| = \sqrt{1+4+4} = 3$ et $\|\mathbf{v}\| = \sqrt{9+0+16} = 5$.

b) $\mathbf{u}\cdot\mathbf{v} = 3+0-8 = -5$, d'où $\cos\theta = -5/15 = -\frac{1}{3}$ et $\theta = \arccos(-\frac{1}{3}) \approx 109{,}47°$. Le produit scalaire est négatif : l'angle est obtus.

c) $\mathbf{v}/\|\mathbf{v}\| = (\frac{3}{5}, 0, -\frac{4}{5})$.

d) Avec $\mathbf{d} = (3,4)$, $\text{proj}_{\mathbf{d}}\, P = \frac{21+4}{25}\,(3,4) = (3,4)$. Le vecteur $P - (3,4) = (4,-3)$ est orthogonal à $\mathbf{d}$, et la distance cherchée est sa norme, $\sqrt{16+9} = 5$.

Exercice D.2.

a) On cherche a, b tels que $a(1,2,1)+b(2,1,0) = (4,5,2)$. La troisième composante donne $a = 2$; la première, $2+2b = 4$, d'où $b = 1$. La deuxième est alors vérifiée : $4+1 = 5$. Donc $\mathbf{w}_1 = 2\mathbf{u} + \mathbf{v}$.

b) La troisième composante impose $a = 1$, la première $1 + 2b = 1$, donc $b = 0$; mais la deuxième donnerait alors $2 = 1$. Le système est incompatible.

c) $\mathbf{u}$ et $\mathbf{v}$ ne sont pas parallèles ; ils engendrent un plan passant par l'origine de $\mathbb{R}^3$. Le vecteur $\mathbf{w}_1$ est dans ce plan, tandis que $\mathbf{w}_2$ en sort.

Exercice D.3.

a) $AB = \begin{pmatrix} 5 & 2 \\ 2 & -2 \end{pmatrix}$ est 2×2, et $BA = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 5 & -1 \\ 0 & 3 & -3 \end{pmatrix}$ est 3×3. Les deux produits n'ont même pas la même taille.

b) $B^\mathsf{T} = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 3 \end{pmatrix}$ et $A^\mathsf{T} = \begin{pmatrix} 1 & 0 \\ 2 & 1 \\ 0 & -1 \end{pmatrix}$, d'où $B^\mathsf{T}A^\mathsf{T} = \begin{pmatrix} 5 & 2 \\ 2 & -2 \end{pmatrix}$, qui est bien la transposée de AB (cette matrice est d'ailleurs symétrique).

c) $R^2 = -I_2$ et $R^4 = I_2$. La matrice R est la rotation d'un quart de tour ; deux quarts de tour font un demi-tour (l'homothétie de rapport -1), et quatre quarts de tour ramènent chaque point à sa place.

Exercice D.4.

$$\left(\begin{array}{ccc|c} 1 & 2 & 1 & 1 \\ 2 & 5 & 3 & 2 \\ -1 & 1 & 4 & 1 \end{array}\right) \xrightarrow[L_3 \leftarrow L_3 + L_1]{L_2 \leftarrow L_2 - 2L_1} \left(\begin{array}{ccc|c} 1 & 2 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 3 & 5 & 2 \end{array}\right) \xrightarrow[L_3 \leftarrow L_3 - 3L_2]{L_1 \leftarrow L_1 - 2L_2} \left(\begin{array}{ccc|c} 1 & 0 & -1 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 2 & 2 \end{array}\right)$$

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

La solution unique est $(x, y, z) = (2, -1, 1)$.

Exercice D.5.

a) Les opérations $L_2 \leftarrow L_2 - L_1$, $L_3 \leftarrow L_3 - L_1$, puis $L_3 \leftarrow L_3 - 2L_2$ donnent

$$\left(\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & k-5 & h-3 \end{array}\right).$$

Si $k \neq 5$, il y a trois pivots : solution unique, quel que soit h. Si $k = 5$ et $h \neq 3$, la dernière ligne s'écrit $0 = h - 3 \neq 0$: aucune solution. Si $k = 5$ et $h = 3$, la dernière ligne est nulle et z est libre : infinité de solutions.

b) Pour $k = 5$, $h = 3$: $z = t$, $y = 1 - 2t$ et $x = 1 - y - z = t$, soit $(x, y, z) = (0, 1, 0) + t\,(1, -2, 1)$. C'est une droite de $\mathbb{R}^3$: les trois plans se coupent selon une même droite (« livre ouvert »). Lorsque $k = 5$ et $h \neq 3$, les plans se coupent deux à deux selon trois droites parallèles, sans point commun (le « prisme »).

Exercice D.6.

a) Dépendante. En échelonnant la matrice dont ce sont les colonnes, la troisième colonne n'a pas de pivot ; on trouve $(1, 2, 3) - 2\,(4, 5, 6) + (7, 8, 9) = \mathbf{0}$.

b) Indépendante : le déterminant de la matrice dont ce sont les colonnes vaut $2 \neq 0$ (proposition D.1).

c) Dépendante, sans aucun calcul : on ne peut pas trouver quatre vecteurs indépendants dans $\mathbb{R}^3$, puisqu'une matrice 3×4 a au plus trois pivots. Une relation se trouve en résolvant le système homogène (par exemple avec Python).

Exercice D.7.

a) Les deux vecteurs ne sont pas parallèles (le déterminant vaut $-2 \neq 0$) ; deux vecteurs indépendants de $\mathbb{R}^2$ forment une base. On résout $a + b = 3$, $a - b = 5$: $a = 4$, $b = -1$. Par produits scalaires, puisque $\mathbf{b}_1 \cdot \mathbf{b}_2 = 0$: $a = \frac{(3,5)\cdot\mathbf{b}_1}{\mathbf{b}_1\cdot\mathbf{b}_1} = \frac{8}{2} = 4$ et $b = \frac{(3,5)\cdot\mathbf{b}_2}{\mathbf{b}_2\cdot\mathbf{b}_2} = \frac{-2}{2} = -1$. Ainsi $(3, 5) = 4\mathbf{b}_1 - \mathbf{b}_2$.

b) L'équation donne $x = y - 2z$, avec y et z libres : $(x, y, z) = y\,(1, 1, 0) + z\,(-2, 0, 1)$. Les vecteurs $(1, 1, 0)$ et $(-2, 0, 1)$ forment une base du plan, qui est de dimension 2.

Exercice D.8.

a) La forme échelonnée réduite de A est $\begin{pmatrix}1&2&0\\0&0&1\\0&0&0\end{pmatrix}$: pivots dans les colonnes 1 et 3, y libre. Le noyau est engendré par $(-2,1,0)$; l'image par les colonnes 1 et 3 de A, soit $(1,2,3)$ et $(-1,1,0)$. On a bien $\dim\ker A + \operatorname{rang} A = 1+2 = 3$.

b) Les deux vecteurs de base vérifient $x+y-z=0$, et ils engendrent un plan : c'est donc ce plan. Le vecteur $(1,1,2)$ vérifie l'équation ; il est dans l'image. Le vecteur $(1,0,0)$ ne la vérifie pas.

c) Une solution particulière est $\mathbf{x}_p = (\frac{2}{3},0,-\frac{1}{3})$. L'ensemble des solutions est $(\frac{2}{3},0,-\frac{1}{3}) + t\,(-2,1,0)$, $t\in\mathbb{R}$: une droite parallèle au noyau.

Exercice D.9.

a) En développant, $\det M_k = k(k^2-1)-(k-1)+(1-k) = k^3-3k+2 = (k-1)^2(k+2)$. La matrice est inversible si et seulement si $k\neq 1$ et $k\neq -2$.

b) Les trois colonnes de M_1 sont égales à $(1,1,1)$: l'image est la droite engendrée par ce vecteur. Le parallélépipède est aplati sur une droite ; son volume est nul.

c) La matrice formée de ces colonnes est triangulaire supérieure, de diagonale $2,3,4$: le volume vaut 24.

d) $\det(2A) = 2^3\cdot 5 = 40$ (chacune des trois lignes est multipliée par 2) ; $\det(A^{-1}) = \frac{1}{5}$, puisque $\det(A)\ \det(A^{-1}) = \det(I) = 1$; et $\det(A^{\mathsf{T}}A) = 5\cdot 5 = 25$.

Exercice D.10.

a) $\det A = 9-8 = 1$, donc $A^{-1} = \begin{pmatrix}3&-2\\-4&3\end{pmatrix}$, et $\mathbf{x} = A^{-1}\begin{pmatrix}1\\2\end{pmatrix} = \begin{pmatrix}-1\\2\end{pmatrix}$.

b) $\det = a^2-4$: la matrice est inversible si et seulement si $a\neq\pm 2$.

c) $R_\theta = \begin{pmatrix}\cos\theta&-\sin\theta\\\sin\theta&\cos\theta\end{pmatrix}$ a pour déterminant $\cos^2\theta+\sin^2\theta = 1$; la formule (D.4) donne $R_\theta^{-1} = \begin{pmatrix}\cos\theta&\sin\theta\\-\sin\theta&\cos\theta\end{pmatrix}$, qui est à la fois la transposée de R_θ et, puisque le cosinus est pair et le sinus impair, la matrice $R_{-\theta}$. Pour défaire une rotation, on tourne du même angle dans l'autre sens ; c'est l'exemple type d'une matrice orthogonale.

Exercice D.11.

a) Les vecteurs de l'axe, multiples de $(1,1)$, sont fixes : valeur propre 1. Les vecteurs perpendiculaires, multiples de $(1,-1)$, sont retournés : valeur propre -1. Par le calcul, $\lambda^2-1=0$, et $\begin{pmatrix}0&1\\1&0\end{pmatrix}\begin{pmatrix}1\\-1\end{pmatrix} = \begin{pmatrix}-1\\1\end{pmatrix}$.

b) $\det(A-\lambda I) = \lambda^2-7\lambda+10 = (\lambda-2)(\lambda-5)$. Pour $\lambda=5$, $A-5I = \begin{pmatrix}-1&1\\2&-2\end{pmatrix}$ a pour noyau $\operatorname{span}\{(1,1)\}$; pour $\lambda=2$, $A-2I = \begin{pmatrix}2&1\\2&1\end{pmatrix}$ a pour noyau $\operatorname{span}\{(1,-2)\}$. Vérification : $5+2 = 7 = 4+3$ (la trace) et $5\cdot 2 = 10 = 12-2$ (le déterminant).

c) Le polynôme caractéristique est $(1-\lambda)^2$: seule valeur propre 1. Le noyau de $\begin{pmatrix}0&1\\0&0\end{pmatrix}$ est $\operatorname{span}\{(1,0)\}$, une seule droite.

Exercice D.12.

a) Avec $P = \begin{pmatrix} 1 & 1 \\ 1 & -2 \end{pmatrix}$ et $D = \begin{pmatrix} 5 & 0 \\ 0 & 2 \end{pmatrix}$, on a $P^{-1} = \frac{1}{3}\begin{pmatrix} 2 & 1 \\ 1 & -1 \end{pmatrix}$ et $A = PDP^{-1}$. Donc

$$A^k = PD^kP^{-1} = \frac{1}{3}\begin{pmatrix} 2\cdot 5^k + 2^k & 5^k - 2^k \\ 2\cdot 5^k - 2\cdot 2^k & 5^k + 2\cdot 2^k \end{pmatrix};$$

pour $k = 1$, on retrouve bien A.

b) Le polynôme caractéristique est $\lambda^2 - 25$: valeurs propres 5 et -5, de vecteurs propres $(2,1)$ et $(1,-2)$, orthogonaux. En les normalisant, $Q = \frac{1}{\sqrt{5}}\begin{pmatrix} 2 & 1 \\ 1 & -2 \end{pmatrix}$ et $D = \begin{pmatrix} 5 & 0 \\ 0 & -5 \end{pmatrix}$. La matrice $\frac{1}{5}S$ a pour valeurs propres 1 et -1, avec des directions propres perpendiculaires : elle fixe la droite engendrée par $(2,1)$ et retourne la direction perpendiculaire. C'est la réflexion d'axe $y = x/2$.

Exercice D.13.

a) `np.linalg.det(A)` renvoie `-125.00000000000004`, soit -125 à un arrondi près. Le déterminant n'étant pas nul, le système a une solution unique (théorème D.2).

b) `x = np.linalg.solve(A, b)` donne `[ 1. -1.  2.  3.]`, et `A @ x - b` renvoie un vecteur nul (ou de l'ordre de 10^{-15}, selon la machine).

c) Les trois méthodes donnent $\mathbf{x} = (1,-1,2,3)$; la forme échelonnée réduite exacte de $(A \mid \mathbf{b})$ est $(I_4 \mid \mathbf{x})$. Pour une matrice aussi petite, les différences d'arrondi entre `solve` et `inv` sont invisibles ; elles deviennent sensibles pour de grandes matrices ou des matrices proches d'être singulières (remarque D.2).

Exercice D.14.

a) La commande `A.row_join(b).rref()` renvoie la matrice

$$\left(\begin{array}{cccc|c} 1 & 2 & 0 & 0 & 1 \\ 0 & 0 & 1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right)$$

et les pivots `(0, 2)` : les pivots sont dans les colonnes de x_1 et de x_3, et x_2 et x_4 sont libres. La troisième équation était redondante ($L_3 = L_2 - L_1$).

b) `A.nullspace()` renvoie les vecteurs $(-2,1,0,0)$ et $(0,0,1,1)$, et `A.rank()` renvoie 2 ; on a bien $2 + 2 = 4$.

c) En posant $x_2 = s$ et $x_4 = t$: $x_1 = 1 - 2s$ et $x_3 = t - 1$, soit

$$\mathbf{x} = (1,0,-1,0) + s\,(-2,1,0,0) + t\,(0,0,1,1), \qquad s,t \in \mathbb{R}\,.$$

C'est un plan (de dimension 2) dans $\mathbb{R}^4$, parallèle au noyau et passant par $(1,0,-1,0)$. La commande `sp.linsolve((A, b))` fournit directement cette description paramétrique.

Exercice D.15. Plusieurs réponses sont possibles ; en voici une.

a) La méthode de réduction : on additionne à une équation un multiple de l'autre afin d'éliminer une inconnue, ce qui est exactement l'opération élémentaire $L_i \leftarrow L_i + c\,L_j$. L'enseignant qui le sait peut justifier la méthode (une opération élémentaire est inversible, donc ne modifie pas l'ensemble des solutions) et répondre aux élèves qui se demandent pourquoi on a « le droit » d'additionner des équations. Il peut aussi présenter la réduction comme la méthode qui se généralise le mieux à trois équations ou plus, contrairement à la comparaison.

b) Chaque équation représente une droite. Des droites sécantes donnent une solution unique ; des droites parallèles distinctes, aucune ; des droites confondues, une infinité. Les deux derniers cas sont exactement ceux où les droites ont la même pente, c'est-à-dire où le déterminant de la matrice des coefficients est nul (théorème D.2). L'enseignant peut ainsi construire volontairement des exemples des trois types, au lieu de ne proposer que des systèmes à solution unique, et amener les élèves à prévoir la nature des solutions en comparant les coefficients avant de calculer.

c) Le travail effectué par une force : pour tirer un traîneau à l'aide d'une corde inclinée, seule la composante de la force parallèle au déplacement est utile, et le travail vaut $\mathbf{F} \cdot \mathbf{d} = \|\mathbf{F}\| \, \|\mathbf{d}\| \cos\theta$. On peut aussi utiliser le produit scalaire pour vérifier qu'un quadrilatère tracé dans un plan cartésien est un rectangle, en montrant que ses côtés consécutifs sont perpendiculaires. Ces deux situations font le lien avec la physique et avec la géométrie analytique de quatrième secondaire.

Annexe E – Éléments de théorie des nombres

Exercice E.1.

a) Si $a \mid b$, il existe $k \in \mathbb{Z}$ tel que $b = ka$. Alors $bc = kac = (kc)\, a$, où $kc \in \mathbb{Z}$. Donc $a \mid bc$.

b) Il existe $k, j \in \mathbb{Z}$ tels que $b = ka$ et $c = jb$. Alors $c = jka = (jk)\, a$, avec $jk \in \mathbb{Z}$: a divise c.

c) Il existe $k, j \in \mathbb{Z}$ tels que $b = ka$ et $c = ja$. Pour tous $x, y \in \mathbb{Z}$, $bx + cy = kax + jay = (kx + jy)\, a$, avec $kx + jy \in \mathbb{Z}$. Donc $a \mid (bx + cy)$.

d) Il existe $k, j \in \mathbb{Z}$ tels que $b = ka$ et $a = jb$. Alors $b = jkb$, et puisque $b \neq 0$, $jk = 1$. Les seuls entiers dont le produit vaut 1 sont $j = k = 1$ et $j = k = -1$. Dans le premier cas $a = b$, dans le second $a = -b$.

e) Il existe $k \in \mathbb{Z}$ tel que $b = ka$. Alors $bc = k\,(ac)$. Comme $a \neq 0$ et $c \neq 0$, on a $ac \neq 0$, et cette égalité dit exactement que $ac \mid bc$.

Exercice E.2.

a) $250 = 13 \times 19 + 3$: le quotient est 19 et le reste 3. Pour -250, on a $-250 = 13 \times (-19) - 3$, qui ne convient pas (reste négatif) ; on descend d'un multiple de 13 : $-250 = 13 \times (-20) + 10$, avec $0 \leq 10 < 13$. Le quotient est -20 et le reste 10.

b) Tout entier n s'écrit $n = 2m$ ou $n = 2m + 1$ (division euclidienne par 2). Dans le premier cas, $n^2 = 4m^2$, qui est divisible par 4. Dans le second, $n^2 = 4m^2 + 4m + 1 = 4(m^2 + m) + 1$, dont le reste modulo 4 est 1.

c) Par la partie b), a^2 et b^2 sont chacun congrus à 0 ou à 1 modulo 4. Leur somme est donc congrue à 0, 1 ou 2 modulo 4, mais jamais à 3. Un entier de la forme $4k + 3$ ne peut donc pas s'écrire $a^2 + b^2$. Par exemple, 3, 7, 11 et 15 ne sont pas des sommes de deux carrés.

Exercice E.3.

a) On a $14 < \sqrt{221} < 15$ et $14 < \sqrt{223} < 15$. Il suffit de tester les nombres premiers $2, 3, 5, 7, 11, 13$, soit six divisions au plus. Pour 221, on trouve $221 = 13 \times 17$: il est composé. Pour 223, aucun de ces six nombres ne divise 223 (les restes sont $1, 1, 3, 6, 3$ et 2) ; par le point (ii) du lemme, 223 est premier.

b) On cherche le plus petit diviseur premier de 6409 parmi les nombres premiers inférieurs à $\sqrt{6409} \approx 80{,}06$. Ni 2, ni 3 (somme des chiffres 19), ni 5, ni 7, ni 11 ne conviennent, mais $6409 = 13 \times 493$. On reprend avec 493 : son plus petit facteur premier est au moins 13 (sinon il diviserait aussi 6409), et au plus $\sqrt{493} \approx 22{,}2$. On teste 13, 17 et 19 : $493 = 17 \times 29$. Enfin, 29 est premier. Ainsi $6409 = 13 \times 17 \times 29$.

c) Pour $2 \leq k \leq n$, l'entier k divise $n! = 1 \times 2 \times \cdots \times n$, et il divise k ; il divise donc la somme $n! + k$. Comme $1 < k < n! + k$, cet entier est composé. On obtient ainsi $n - 1$ entiers consécutifs sans aucun nombre premier. Puisque n est arbitraire, il existe des écarts aussi grands qu'on veut entre deux nombres premiers consécutifs. Par exemple, avec $n = 5$, les entiers $122, 123, 124, 125$ sont composés.

Exercice E.4.

a) Un entier est congru à 0, 1, 2 ou 3 modulo 4. S'il est congru à 0 ou à 2, il est pair. Un nombre premier impair est donc congru à 1 ou à 3 modulo 4.

b) Si $a \equiv 1$ et $b \equiv 1 \pmod 4$, alors $ab \equiv 1 \times 1 = 1 \pmod 4$ par la règle du produit. Par récurrence, un produit de k facteurs congrus à 1 est congru à 1.

c) On a $N \equiv -1 \equiv 3 \pmod 4$ et $N \geq 4 \times 3 - 1 = 11$. Comme N est impair, tous ses facteurs premiers sont impairs, donc congrus à 1 ou à 3 modulo 4 par a). S'ils étaient tous congrus à 1, leur produit N serait congru à 1 par b), ce qui n'est pas le cas. Donc N a un facteur premier $p \equiv 3 \pmod 4$. Par hypothèse, p est l'un des p_i. Alors p divise $4p_1 \cdots p_r$ et N, donc leur différence, qui vaut 1 : c'est impossible. Il y a donc une infinité de nombres premiers de la forme $4k + 3$.

d) Si l'on part de nombres premiers $p_i \equiv 1 \pmod 4$ et qu'on forme $N = 4p_1 \cdots p_r + 1 \equiv 1 \pmod 4$, rien n'oblige N à avoir un facteur premier congru à 1 : un produit de deux nombres congrus à 3 est aussi congru à 1 ($3 \times 3 = 9 \equiv 1$). Le résultat reste vrai pour les nombres premiers de la forme $4k + 1$, mais sa démonstration demande d'autres outils.

Exercice E.5.

a) Les divisions successives sont

$$\begin{aligned} 42\,823 &= 6 \times 6409 + 4369, \\ 6409 &= 1 \times 4369 + 2040, \\ 4369 &= 2 \times 2040 + 289, \\ 2040 &= 7 \times 289 + 17, \\ 289 &= 17 \times 17 + 0. \end{aligned}$$

Le dernier reste non nul est 17 : $\operatorname{pgcd}(42\,823, 6409) = 17$.

b) On remonte les divisions, en partant de l'avant-dernière :

$$\begin{aligned} 17 &= 2040 - 7 \times 289 \\ &= 2040 - 7\,(4369 - 2 \times 2040) = 15 \times 2040 - 7 \times 4369 \\ &= 15\,(6409 - 4369) - 7 \times 4369 = 15 \times 6409 - 22 \times 4369 \\ &= 15 \times 6409 - 22\,(42\,823 - 6 \times 6409) = 147 \times 6409 - 22 \times 42\,823. \end{aligned}$$

On peut donc prendre $x = -22$ et $y = 147$. Ce n'est pas la seule possibilité : par la proposition E.16, les solutions sont $x = -22 + 377t$ et $y = 147 - 2519t$, où $377 = 6409/17$ et $2519 = 42\,823/17$.

c) On a vu que $6409 = 13 \times 17 \times 29$. Puisque 17 divise 42 823 par a), on calcule $42\,823 = 17 \times 2519$. Le plus petit facteur premier de 2519 est au plus $\sqrt{2519} \approx 50{,}2$; on trouve $2519 = 11 \times 229$. Enfin, 229 est premier (aucun des nombres premiers $2, 3, 5, 7, 11, 13$, inférieurs à $\sqrt{229} \approx 15{,}1$, ne le divise). Ainsi $42\,823 = 11 \times 17 \times 229$.

d) Le seul nombre premier commun aux deux factorisations est 17, avec l'exposant 1 dans les deux cas. Par la proposition E.15, le PGCD est $17^1 = 17$. On remarque que connaître le PGCD à l'avance facilitait la factorisation : il fournissait directement un facteur commun.

Exercice E.6.

a) En centimètres, le plancher mesure 315×135. Par la proposition E.8, le plus grand côté possible est $\operatorname{pgcd}(315, 135)$. L'algorithme d'Euclide donne $315 = 2 \times 135 + 45$ et $135 = 3 \times 45 + 0$: le PGCD est 45. Les tuiles mesurent donc 45 cm de côté, et il en faut $\frac{315}{45} \times \frac{135}{45} = 7 \times 3 = 21$.

b) On découpe d'abord deux carrés de côté 135 (premier quotient 2), ce qui laisse une bande de 45×135 (premier reste 45). Cette bande se découpe exactement en trois carrés de côté 45 (deuxième quotient 3, reste nul). Le dessin compte donc deux grands carrés côte à côte et une bande verticale formée de trois petits carrés ; le côté des plus petits carrés est le PGCD.

c) Non : 20 ne divise ni 315 ni 135 (leurs restes dans la division par 20 valent 15). Par la proposition E.8, des carrés de côté 20 ne peuvent pas paver le plancher. Plus généralement, les seules tailles entières possibles sont les diviseurs de 45 : 1, 3, 5, 9, 15 et 45 cm.

Exercice E.7.

a) $360 = 2^3 \cdot 3^2 \cdot 5$ et $84 = 2^2 \cdot 3 \cdot 7$. Par la proposition E.15, $\operatorname{pgcd}(360, 84) = 2^2 \cdot 3 = 12$ et $\operatorname{ppcm}(360, 84) = 2^3 \cdot 3^2 \cdot 5 \cdot 7 = 2520$. On a bien $12 \times 2520 = 30\,240 = 360 \times 84$.

b) Puisque 12 divise a et b, écrivons $a = 12a'$ et $b = 12b'$. Les entiers a' et b' sont premiers entre eux, sinon un diviseur commun $e > 1$ ferait de $12e$ un diviseur commun de a et b plus grand que 12. Par la proposition E.15, $ab = 12 \times 360$, soit $144\,a'b' = 4320$, d'où $a'b' = 30$. Il faut donc écrire $30 = 2 \cdot 3 \cdot 5$ comme produit de deux facteurs premiers entre eux, $a' \leq b'$; chaque nombre premier va entièrement dans l'un ou l'autre facteur. Les possibilités sont $(1, 30)$, $(2, 15)$, $(3, 10)$ et $(5, 6)$. Les couples cherchés sont $(12, 360)$, $(24, 180)$, $(36, 120)$ et $(60, 72)$. On vérifie par exemple que $60 = 2^2 \cdot 3 \cdot 5$ et $72 = 2^3 \cdot 3^2$ ont pour PGCD $2^2 \cdot 3 = 12$ et pour PPCM $2^3 \cdot 3^2 \cdot 5 = 360$.

Exercice E.8.

a) On a $(n+1) \cdot 1 + n \cdot (-1) = 1$, et $3\,(2n+1) - 2\,(3n+1) = 6n + 3 - 6n - 2 = 1$. Par le point (iii) du corollaire E.10, les deux paires sont premières entre elles.

b) Non. On a $\operatorname{pgcd}(15, 21) = 3$, et 3 divise $15x + 21y$ pour tous les entiers x et y ; il ne peut donc pas diviser 10.

c) Puisque $3 \mid 9$, il y a des solutions. En divisant par 3, l'équation devient $5x + 7y = 3$. On a $5 \times 3 + 7 \times (-2) = 1$ (ou, par l'algorithme d'Euclide, $7 = 1 \times 5 + 2$, $5 = 2 \times 2 + 1$, d'où $1 = 5 - 2 \times 2 = 5 - 2(7 - 5) = 3 \times 5 - 2 \times 7$). En multipliant par 3, on obtient la solution particulière $(x_0, y_0) = (9, -6)$. Par la proposition E.16, avec $b/d = 7$ et $a/d = 5$, les solutions sont

$$x = 9 + 7t, \qquad y = -6 - 5t, \qquad t \in \mathbb{Z}\,.$$

Avec $t = -1$, par exemple, $(x, y) = (2, -1) : 30 - 21 = 9$.

d) L'égalité dit qu'en versant trois fois 5 L dans un récipient et en en retirant deux fois 7 L, il reste 1 L. Concrètement : on remplit le récipient de 5 L et on le vide dans celui de 7 L ; on le remplit de nouveau et on complète le récipient de 7 L, ce qui laisse 3 L dans celui de 5 L ; on vide le récipient de 7 L dans l'évier et on y verse les 3 L ; on remplit une troisième fois le récipient de 5 L et on complète celui de 7 L (il manque 4 L), ce qui laisse exactement 1 L dans le récipient de 5 L. On a bien rempli trois fois le récipient de 5 L et vidé deux fois le récipient de 7 L.

Exercice E.9.

a) Si $a \equiv b \pmod m$, alors $m \mid a - b$ par définition. Par le point (ii) de la proposition E.1, m divise $(-1)(a - b) = b - a$, c'est-à-dire $b \equiv a \pmod m$. Le même raisonnement, en échangeant a et b, donne la réciproque. Par ailleurs, $a - b \equiv 0 \pmod m$ signifie que m divise $(a - b) - 0 = a - b$, ce qui est exactement la définition de $a \equiv b \pmod m$.

b) Par hypothèse, m divise $a-b$ et $c-d$. Il divise donc leur somme $(a-b)+(c-d) = (a + c) - (b + d)$, d'où $a + c \equiv b + d \pmod m$.

c) Par hypothèse, m divise $a-b$ et $b-c$, donc aussi leur somme, $(a-b)+(b-c) = a - c$. Ainsi $a \equiv c \pmod m$.

d) Par hypothèse, m divise $ax - ay = a(x - y)$. Puisque $\operatorname{pgcd}(a, m) = 1$, le lemme de Gauss donne $m \mid x - y$, d'où $x \equiv y \pmod m$.

e) Prenons $a = 2$, $x = 5$, $y = 0$ et $m = 10$. On a $ax = 10 \equiv 0 = ay \pmod{10}$, mais $5 \not\equiv 0 \pmod{10}$. L'hypothèse de d) n'est pas satisfaite, puisque $\operatorname{pgcd}(2, 10) = 2$.

Exercice E.10.

a) Comme $10 \equiv -1 \pmod{11}$, on a $10^k \equiv (-1)^k \pmod{11}$ par le point (iii) de la proposition E.18. Par les règles de la somme et du produit,

$$\overline{c_m \cdots c_0} = \sum_{k=0}^{m} c_k \, 10^k \equiv \sum_{k=0}^{m} (-1)^k c_k \pmod{11}.$$

Un entier est donc divisible par 11 si et seulement si la somme alternée de ses chiffres, en partant des unités, est divisible par 11.

b) Pour 123 456 789, la somme des chiffres vaut 45, qui est divisible par 9 : le nombre est divisible par 9. La somme alternée vaut $9-8+7-6+5-4+3-2+1 = 5$: il n'est pas divisible par 11 (son reste modulo 11 est 5). Pour 918 082, la somme des chiffres vaut $28 \equiv 1 \pmod 9$: il n'est pas divisible par 9. La somme alternée vaut $2 - 8 + 0 - 8 + 1 - 9 = -22$, divisible par 11 : le nombre est divisible par 11 (en effet, $918\,082 = 11 \times 83\,462$).

c) La somme alternée des chiffres de $\overline{abba}$ vaut $a - b + b - a = 0$, qui est divisible par 11. Directement, $\overline{abba} = 1001a + 110b = 11\,(91a + 10b)$.

Exercice E.11.

a) On calcule $S = 1 \cdot 2 + 2 \cdot 7 + 3 \cdot 6 + 4 \cdot 5 + 5 \cdot 4 + 6 \cdot 0 + 7 \cdot 9 + 8 \cdot 1 + 9 \cdot 2 = 2+14+18+20+20+0+63+8+18 = 163$. La condition s'écrit $163+10\,x_{10} \equiv 0 \pmod{11}$. Comme $163 = 14 \times 11 + 9 \equiv 9$ et $10 \equiv -1$, elle devient $9 - x_{10} \equiv 0$, soit $x_{10} = 9$. Le code est 2 7 6 5 4 0 9 1 2 9.

b) Si l'on remplace le chiffre x_i par $y_i \neq x_i$, la somme pondérée change de $i\,(y_i - x_i)$. Pour que le nouveau code soit valide, il faudrait que $11 \mid i\,(y_i - x_i)$. Par le lemme

d'Euclide (11 est premier), cela exigerait $11 \mid i$ ou $11 \mid y_i - x_i$. Or $1 \leq i \leq 10$ et $0 < |y_i - x_i| \leq 10$: aucun de ces deux nombres n'est divisible par 11. L'erreur est donc toujours détectée.

c) Si l'on échange x_i et x_{i+1}, la somme pondérée change de

$$(i\,x_{i+1} + (i+1)\,x_i) - (i\,x_i + (i+1)\,x_{i+1}) = x_i - x_{i+1}.$$

Comme $0 < |x_i - x_{i+1}| \leq 10$, cette variation n'est pas divisible par 11, et l'erreur est détectée.

d) Le raisonnement de b) repose sur le fait que 11 est premier. Modulo 10, il échouerait : remplacer $x_2 = 0$ par $y_2 = 5$ change la somme de $2 \times 5 = 10 \equiv 0 \pmod{10}$, et l'erreur passerait inaperçue. Le module premier 11 garantit que tous les poids $1, \ldots, 10$ sont inversibles. Le prix à payer est un onzième symbole, X, pour le chiffre de contrôle.

Exercice E.12.

a) Comme $15 \equiv -1 \pmod 4$, la congruence devient $-x \equiv 2$, soit $x \equiv -2 \equiv 2 \pmod 4$. On vérifie : $15 \times 2 = 30 = 7 \times 4 + 2$.

b) $12x \equiv 0 \pmod 3$ pour tout x, alors que $4 \equiv 1 \pmod 3$. Il n'y a aucune solution, ce que confirme la proposition E.20 : $\operatorname{pgcd}(12, 3) = 3$ ne divise pas 4.

c) $\operatorname{pgcd}(2, 6) = 2$ ne divise pas 3 : aucune solution. En effet, $2x - 3$ est impair et ne peut pas être un multiple de 6.

d) $\operatorname{pgcd}(2, 7) = 1$, il y a une solution unique modulo 7. L'inverse de 2 modulo 7 est 4 ($2 \times 4 = 8 \equiv 1$). En multipliant par 4, $x \equiv 12 \equiv 5 \pmod 7$. On vérifie : $2 \times 5 = 10 \equiv 3 \pmod 7$.

e) $\operatorname{pgcd}(6, 10) = 2$ divise 4 : il y a deux solutions modulo 10. En divisant par 2, on obtient $3x \equiv 2 \pmod 5$. L'inverse de 3 modulo 5 est 2, d'où $x \equiv 4 \pmod 5$. Les solutions modulo 10 sont $x \equiv 4$ et $x \equiv 9 \pmod{10}$ ($6 \times 4 = 24$ et $6 \times 9 = 54$ sont bien congrus à 4 modulo 10).

Exercice E.13.

a) On remarque que $12 \times 2 = 24 = 23 + 1 \equiv 1 \pmod{23}$. L'inverse est 2. (L'algorithme d'Euclide donne la même chose : $23 = 1 \times 12 + 11$, $12 = 1 \times 11 + 1$, d'où $1 = 12 - 11 = 12 - (23 - 12) = 2 \times 12 - 23$.)

b) $\operatorname{pgcd}(33, 114) = 3$, puisque $33 = 3 \times 11$ et $114 = 2 \times 3 \times 19$. Par la proposition E.19, 33 n'est pas inversible modulo 114.

c) $101 \equiv -1 \pmod{102}$, et $(-1)(-1) = 1$. Donc 101 est son propre inverse : $101^2 = 10\,201 = 100 \times 102 + 1$.

d) $43 = 2 \times 17 + 9$, $17 = 1 \times 9 + 8$, $9 = 1 \times 8 + 1$. En remontant,

$$1 = 9 - 8 = 9 - (17 - 9) = 2 \times 9 - 17 = 2\,(43 - 2 \times 17) - 17 = 2 \times 43 - 5 \times 17.$$

Ainsi $17 \times (-5) \equiv 1 \pmod{43}$, et l'inverse est $-5 \equiv 38$. On vérifie : $17 \times 38 = 646 = 15 \times 43 + 1$.

Exercice E.14.

a) 7 est premier et ne divise pas 3, donc $3^6 \equiv 1 \pmod 7$. Comme $100 = 16 \times 6 + 4$, $3^{100} = (3^6)^{16} \cdot 3^4 \equiv 3^4 = 81 = 11 \times 7 + 4 \equiv 4 \pmod 7$. Le reste est 4.

b) La somme compte $p - 2$ termes, de 2^{p-1} à $(p-1)^{p-1}$. Chaque base k vérifie $2 \leq k \leq p - 1$, donc $p \nmid k$, et le petit théorème de Fermat donne $k^{p-1} \equiv 1$

$\pmod p$. La somme est donc congrue à $1+1+\cdots+1 = p-2 \equiv -2 \pmod p$. (Pour $p=2$, la somme est vide et vaut $0 \equiv -2 \pmod 2$.) Par exemple, pour $p=5$: $2^4+3^4+4^4 = 16+81+256 = 353 = 71\times 5-2$.

c) $2^{10} = 1024 = 3\times 341+1$, donc $2^{10} \equiv 1 \pmod{341}$. Alors $2^{340} = (2^{10})^{34} \equiv 1^{34} = 1 \pmod{341}$. Le nombre $n=341$ vérifie donc $2^{n-1} \equiv 1 \pmod n$, comme le ferait un nombre premier, et pourtant il est composé. La réciproque du petit théorème de Fermat est fausse : la congruence $2^{n-1}\equiv 1 \pmod n$ ne suffit pas à prouver que n est premier. On dit que 341 est un nombre pseudo-premier en base 2.

Exercice E.15.

a) Par le corollaire E.22, appliqué à $p=5$, $n^5 \equiv n \pmod 5$, donc $5 \mid n^5-n$. Pour 2 et 3, on factorise :

$$n^5-n = n(n^4-1) = n(n^2-1)(n^2+1) = (n-1)\,n\,(n+1)\,(n^2+1).$$

Le produit $(n-1)n(n+1)$ de trois entiers consécutifs contient un multiple de 2 et un multiple de 3 ; il est donc divisible par 2 et par 3. (On peut aussi appliquer le corollaire E.22 avec $p=2$ et $p=3$.) Ainsi 2, 3 et 5 divisent n^5-n. Comme ils sont premiers entre eux deux à deux, le corollaire E.13, appliqué deux fois, donne $2\times 3\times 5 = 30 \mid n^5-n$.

b) On a $n^2(n^4-1) = n\,(n^5-n)$, qui est divisible par 3 et par 5 par la partie a). Montrons qu'il est divisible par 4. Si n est pair, $n=2k$ et $n^2=4k^2$ est divisible par 4. Si n est impair, $n-1$ et $n+1$ sont deux nombres pairs, donc $(n-1)(n+1)=n^2-1$ est divisible par 4, et $n^4-1=(n^2-1)(n^2+1)$ aussi. Dans les deux cas, $4 \mid n^2(n^4-1)$. Comme 3, 4 et 5 sont premiers entre eux deux à deux, on conclut que $60 \mid n^2(n^4-1)$. Il ne suffisait pas de dire que n^2 est pair : il ne l'est que si n l'est. Par exemple, pour $n=3$, $9\times 80 = 720 = 12\times 60$.

Exercice E.16.

a) La valeur $\varphi(p^k)$ est le nombre d'entiers de 1 à p^k premiers avec p^k. Le seul facteur premier de p^k étant p, un entier est premier avec p^k si et seulement s'il n'est pas divisible par p. Les entiers de 1 à p^k divisibles par p sont $1\cdot p, 2\cdot p, 3\cdot p, \ldots, p^{k-1}\cdot p$; il y en a p^{k-1}. Il reste donc $p^k - p^{k-1}$ entiers premiers avec p^k.

b) $12 = 2^2\cdot 3$: $\varphi(12) = 12\cdot \frac12\cdot\frac23 = 4$. $100 = 2^2\cdot 5^2$: $\varphi(100) = 100\cdot\frac12\cdot\frac45 = 40$. $360 = 2^3\cdot 3^2\cdot 5$: $\varphi(360) = 360\cdot\frac12\cdot\frac23\cdot\frac45 = 96$.

c) Les entiers sont 1, 5, 7 et 11, et $\varphi(12)=4$. Leurs carrés sont 1, $25 = 2\times 12+1$, $49 = 4\times 12+1$ et $121 = 10\times 12+1$: tous congrus à 1 modulo 12. Par conséquent, leurs puissances quatrièmes le sont aussi, ce qui vérifie $a^{\varphi(12)} = a^4 \equiv 1 \pmod{12}$. On constate même que l'exposant 2 suffit ici : $\varphi(n)$ est un exposant qui fonctionne toujours, mais pas forcément le plus petit.

Exercice E.17.

a) Soit $n \geq 1$ un entier, et a un entier tel que $\operatorname{pgcd}(a,n) = 1$. Alors $a^{\varphi(n)} \equiv 1 \pmod n$, où $\varphi(n)$ est le nombre d'entiers de 1 à n premiers avec n.

b) Les deux derniers chiffres sont le reste modulo 100. Comme $\operatorname{pgcd}(7,100) = 1$ et $\varphi(100) = 40$, on a $7^{40} \equiv 1 \pmod{100}$. Or $222 = 5\times 40+22$, donc $7^{222} \equiv 7^{22} \pmod{100}$. On peut aller plus vite en remarquant que $7^4 = 2401 \equiv 1 \pmod{100}$; comme $222 = 4\times 55+2$, $7^{222} \equiv 7^2 = 49 \pmod{100}$. Les deux derniers chiffres sont 49.

c) Les trois derniers chiffres sont le reste modulo 1000. On a $\text{pgcd}(43, 1000) = 1$ et $\varphi(1000) = 1000 \cdot \frac{1}{2} \cdot \frac{4}{5} = 400$, donc $43^{400} \equiv 1 \pmod{1000}$. Ainsi $43^{402} = 43^{400} \cdot 43^2 \equiv 43^2 = 1849 \equiv 849 \pmod{1000}$. Les trois derniers chiffres sont 849. Avec Python, on vérifie le résultat par `pow(43, 402, 1000)`, qui calcule directement la puissance modulaire.

Exercice E.18.

a) Les modules 2 et 4 ne sont pas premiers entre eux, et le théorème ne s'applique pas directement. On remarque que la condition $x \equiv 1 \pmod{4}$ entraîne $x \equiv 1 \pmod{2}$: la première congruence est superflue. Il reste $x \equiv 1 \pmod{3}$ et $x \equiv 1 \pmod{4}$, avec $\text{pgcd}(3, 4) = 1$. La solution évidente $x = 1$ est unique modulo 12 : $x \equiv 1 \pmod{12}$. Plus simplement, $x - 1$ doit être un multiple de 2, 3 et 4, donc de leur PPCM, 12.

b) On ramène d'abord chaque congruence à la forme $x \equiv \ldots$. L'inverse de 2 modulo 5 est 3, donc $x \equiv 3 \times 3 = 9 \equiv 4 \pmod{5}$. L'inverse de 3 modulo 8 est 3 ($3 \times 3 = 9 \equiv 1$), donc $x \equiv 3 \pmod{8}$. On écrit $x = 3 + 8k$; alors $3 + 8k \equiv 4 \pmod{5}$, soit $3k \equiv 1 \pmod{5}$, d'où $k \equiv 2 \pmod{5}$ ($3 \times 2 = 6 \equiv 1$). Avec $k = 2 + 5j$, $x = 19 + 40j$: $x \equiv 19 \pmod{40}$. On vérifie : $2 \times 19 = 38 \equiv 3 \pmod{5}$ et $3 \times 19 = 57 \equiv 1 \pmod{8}$.

c) Le nombre x d'œufs vérifie $x - 1$ divisible par 2, 3, 4, 5 et 6, donc par leur PPCM, 60 : $x \equiv 1 \pmod{60}$. Il faut aussi $x \equiv 0 \pmod{7}$. Avec $x = 1 + 60k$, on obtient $1 + 60k \equiv 1 + 4k \equiv 0 \pmod{7}$, soit $4k \equiv -1 \equiv 6 \pmod{7}$. L'inverse de 4 modulo 7 est 2, d'où $k \equiv 12 \equiv 5 \pmod{7}$. Les solutions sont $x = 301 + 420j$, et le plus petit nombre d'œufs est $301 = 43 \times 7$.

Exercice E.19. Plusieurs réponses sont possibles ; en voici une.

a) La factorisation donne $84 = 2^2 \cdot 3 \cdot 7$ et $126 = 2 \cdot 3^2 \cdot 7$, d'où $\text{pgcd} = 2 \cdot 3 \cdot 7 = 42$. Avec l'algorithme d'Euclide, $126 = 1 \times 84 + 42$ et $84 = 2 \times 42$: le PGCD est 42. Sur du papier quadrillé, on découpe dans un rectangle 126×84 un carré 84×84, puis la bande restante 42×84 se découpe en deux carrés de côté 42. La factorisation est naturelle pour les élèves, réinvestit les nombres premiers et donne aussi le PPCM, mais elle devient laborieuse pour des nombres comme 42 823. L'algorithme d'Euclide est rapide et ne demande que des divisions ; sa version géométrique donne du sens au PGCD et rejoint les situations de dallage étudiées en classe. Son inconvénient est que sa justification (lemme E.6) n'est pas évidente pour les élèves, même si le découpage en carrés la rend visible.

b) Une somme de chiffres est congrue au nombre lui-même modulo 9 (exemple E.6). Si $a \times b = c$, la règle du produit donne donc $s(a) \times s(b) \equiv s(c) \pmod{9}$, où s désigne la somme des chiffres. Si cette congruence est fausse, le calcul est certainement erroné. La réciproque est fausse : la preuve par neuf ne voit que les restes modulo 9. Par exemple, pour $23 \times 14 = 322$, le résultat erroné 232 (chiffres permutés) a la même somme de chiffres que 322 et passe le test. De même, une erreur qui modifie le résultat d'un multiple de 9 (comme 331 au lieu de 322) n'est pas détectée.

c) On peut demander aux élèves de calculer $2 + 1$, $2 \times 3 + 1$, $2 \times 3 \times 5 + 1$, $2 \times 3 \times 5 \times 7 + 1$, et de vérifier si ces nombres (3, 7, 31, 211) sont premiers, puis de poursuivre jusqu'à 30 031, qui ne l'est pas. On leur fait ensuite remarquer que la division de ces nombres par 2, par 3, par 5, … laisse toujours un reste 1. La question « si quelqu'un prétendait avoir la liste de tous les nombres premiers, comment le contredire ? » amène alors l'argument d'Euclide. L'activité permet de discuter de ce qu'est une démonstration par l'absurde, et de corriger l'idée

fausse selon laquelle le produit des premiers nombres premiers, augmenté de 1, est toujours premier (remarque E.3).

Bibliographie

[AF91] Stefan Aeberhard et Michele Forina. *Wine*. UCI Machine Learning Repository. https://doi.org/10.24432/C5PC7J. 1991.

[Bac00] Louis Bachelier. “Théorie de la spéculation”. In : *Annales scientifiques de l’École normale supérieure*. 3e sér. 17 (1900), p. 21-86.

[Ban22] Stefan Banach. “Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales”. In : *Fundamenta Mathematicae* 3 (1922), p. 133-181.

[Bar93] Michael F. Barnsley. *Fractals Everywhere*. 2e éd. Boston : Academic Press, 1993.

[BD85] Michael F. Barnsley et Stephen Demko. “Iterated function systems and the global construction of fractals”. In : *Proceedings of the Royal Society of London. Series A* 399.1817 (1985), p. 243-275.

[BL06] Kurt Bryan et Tanya Leise. “The $25,000,000,000 Eigenvector : The Linear Algebra behind Google”. In : *SIAM Review* 48.3 (2006), p. 569-581.

[BP98] Sergey Brin et Lawrence Page. “The Anatomy of a Large-Scale Hypertextual Web Search Engine”. In : *Computer Networks and ISDN Systems* 30.1–7 (1998), p. 107-117.

[Cac30] Renato Caccioppoli. “Un teorema generale sull’esistenza di elementi uniti in una trasformazione funzionale”. In : *Rendiconti della Reale Accademia Nazionale dei Lincei* 11 (1930), p. 794-799.

[Cra05] John J. Craig. *Introduction to Robotics : Mechanics and Control*. 3e éd. Upper Saddle River, NJ : Pearson Prentice Hall, 2005.

[DH76] Whitfield Diffie et Martin E. Hellman. “New Directions in Cryptography”. In : *IEEE Transactions on Information Theory* 22.6 (1976), p. 644-654.

[Edw97] Laurie D. Edwards. “Exploring the territory before proof : Students’ generalizations in a computer microworld for transformation geometry”. In : *International Journal of Computers for Mathematical Learning* 2.3 (1997), p. 187-215.

[Euc94] Euclide. *Les Éléments, volume 2 : livres V à IX*. Traduction et commentaires de Bernard Vitrac. Paris : Presses universitaires de France, 1994.

[Fal14] Kenneth Falconer. *Fractal Geometry : Mathematical Foundations and Applications*. 3e éd. Chichester : Wiley, 2014.

[Fol+96] James D. Foley et al. *Computer Graphics : Principles and Practice*. 2e éd. Reading, MA : Addison-Wesley, 1996.

[Gal86] Francis Galton. “Regression Towards Mediocrity in Hereditary Stature”. In : *The Journal of the Anthropological Institute of Great Britain and Ireland* 15 (1886), p. 246-263.

[Gau01] Carl Friedrich Gauss. *Disquisitiones arithmeticae*. Leipzig : Gerhard Fleischer, 1801.

[Ger08] Jean-Jacques Gervais. “Mathématiques et technologie — notes de cours”. Notes de cours, Département de mathématiques et de statistique, Université

Laval. Voir en particulier l'annexe 1, « Transformations linéaires ou affines du plan, point de vue géométrique ». 2008.

[Gol91] David Goldberg. "What Every Computer Scientist Should Know About Floating-Point Arithmetic". In : *ACM Computing Surveys* 23.1 (1991), p. 5-48.

[Grc11] Joseph F. Grcar. "Mathematicians of Gaussian Elimination". In : *Notices of the American Mathematical Society* 58.6 (2011), p. 782-792.

[GWF14] Kristen B. Gorman, Tony D. Williams et William R. Fraser. "Ecological Sexual Dimorphism and Environmental Variability within a Community of Antarctic Penguins (Genus *Pygoscelis*)". In : *PLoS ONE* 9.3 (2014), e90081.

[Hau14] Felix Hausdorff. *Grundzüge der Mengenlehre*. Leipzig : Veit & Comp., 1914.

[Hay13] Brian Hayes. "First Links in the Markov Chain". In : *American Scientist* 101.2 (2013), p. 92-97.

[HHG20] Allison M. Horst, Alison P. Hill et Kristen B. Gorman. *palmerpenguins : Palmer Archipelago (Antarctica) Penguin Data*. Paquet R, version 0.1.0. https://allisonhorst.github.io/palmerpenguins/. 2020.

[Hil29] Lester S. Hill. "Cryptography in an Algebraic Alphabet". In : *The American Mathematical Monthly* 36.6 (1929), p. 306-312.

[HLP17] François Husson, Sébastien Lê et Jérôme Pagès. *Exploratory Multivariate Analysis by Example Using R*. 2e éd. Boca Raton : CRC Press, 2017.

[HMW+20] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt et al. "Array programming with NumPy". In : *Nature* 585 (2020), p. 357-362. doi : 10.1038/s41586-020-2649-2.

[Hol03] Karen F. Hollebrands. "High school students' understandings of geometric transformations in the context of a technological environment". In : *The Journal of Mathematical Behavior* 22.1 (2003), p. 55-72.

[Hot33] Harold Hotelling. "Analysis of a Complex of Statistical Variables into Principal Components". In : *Journal of Educational Psychology* 24.6 (1933), p. 417-441.

[Hun07] John D. Hunter. "Matplotlib : A 2D graphics environment". In : *Computing in Science & Engineering* 9.3 (2007), p. 90-95. doi : 10.1109/MCSE.2007.55.

[Hut81] John E. Hutchinson. "Fractals and self similarity". In : *Indiana University Mathematics Journal* 30.5 (1981), p. 713-747.

[JC16] Ian T. Jolliffe et Jorge Cadima. "Principal Component Analysis : A Review and Recent Developments". In : *Philosophical Transactions of the Royal Society A* 374.2065 (2016), p. 20150202.

[Jol02] Ian T. Jolliffe. *Principal Component Analysis*. 2e éd. New York : Springer, 2002.

[Ker83] Auguste Kerckhoffs. "La cryptographie militaire". In : *Journal des sciences militaires* 9 (1883), p. 5-38, 161-191.

[Kle93] Felix Klein. "Vergleichende Betrachtungen über neuere geometrische Forschungen". In : *Mathematische Annalen* 43.1 (1893). Texte original de 1872, dit « programme d'Erlangen », p. 63-100.

[Knu97] Donald E. Knuth. *The Art of Computer Programming, volume 2 : Seminumerical Algorithms*. 3e éd. Reading (Massachusetts) : Addison-Wesley, 1997.

[Koc04] Helge von Koch. “Sur une courbe continue sans tangente, obtenue par une construction géométrique élémentaire”. In : *Arkiv för Matematik, Astronomi och Fysik* 1 (1904), p. 681-702.

[Kom97] John Kominek. “Advances in fractal compression for multimedia applications”. In : *Multimedia Systems* 5.4 (1997), p. 255-270.

[Lei03] Gottfried Wilhelm Leibniz. “Explication de l’arithmétique binaire, qui se sert des seuls caractères 0 et 1, avec des remarques sur son utilité, et sur ce qu’elle donne le sens des anciennes figures chinoises de Fohy”. In : *Mémoires de l’Académie royale des sciences* (1703), p. 85-89.

[LM06] Amy N. Langville et Carl D. Meyer. *Google’s PageRank and Beyond : The Science of Search Engine Rankings*. Princeton : Princeton University Press, 2006.

[Man63] Benoît Mandelbrot. “The variation of certain speculative prices”. In : *The Journal of Business* 36.4 (1963), p. 394-419.

[Man67] Benoît Mandelbrot. “How long is the coast of Britain ? Statistical self-similarity and fractional dimension”. In : *Science* 156.3775 (1967), p. 636-638.

[Man75] Benoît Mandelbrot. *Les objets fractals : forme, hasard et dimension*. Paris : Flammarion, 1975.

[Man82] Benoît Mandelbrot. *The Fractal Geometry of Nature*. San Francisco : W. H. Freeman, 1982.

[Man99] Benoît Mandelbrot. “A multifractal walk down Wall Street”. In : *Scientific American* 280.2 (1999), p. 70-73.

[Mar06] Andreï A. Markov. “Rasprostranenie zakona bol’shih chisel na velichiny, zavisyashchie drug ot druga”. In : *Izvestiya Fiziko-matematicheskogo obshchestva pri Kazanskom universitete, 2e série* 15 (1906). En russe. Titre traduit : « Extension de la loi des grands nombres à des grandeurs qui dépendent les unes des autres », p. 135-156.

[Mar13] Andreï A. Markov. “Primer statisticheskogo issledovaniya nad tekstom « Evgeniya Onegina », illyustriruyushchij svyaz’ ispytanij v tsep”’. In : *Bulletin de l’Académie impériale des sciences de Saint-Pétersbourg, 6e série* 7.3 (1913). En russe. Titre traduit : « Exemple d’une étude statistique du texte d’*Eugène Onéguine*, illustrant la liaison des épreuves en chaîne », p. 153-162.

[MH04] Benoît Mandelbrot et Richard L. Hudson. *The (Mis)behavior of Markets : A Fractal View of Risk, Ruin, and Reward*. New York : Basic Books, 2004.

[Min07] Ministère de l’Éducation, du Loisir et du Sport. *Programme de formation de l’école québécoise — Enseignement secondaire, deuxième cycle. Domaine de la mathématique, de la science et de la technologie*. Gouvernement du Québec, Québec. 2007.

[Min16] Ministère de l’Éducation et de l’Enseignement supérieur. *Progression des apprentissages au secondaire — Mathématique*. Gouvernement du Québec, Québec. 2016.

[Mor46] Patrick A. P. Moran. “Additive functions of intervals and Hausdorff measure”. In : *Proceedings of the Cambridge Philosophical Society* 42.1 (1946), p. 15-23.

[MSP+17] Aaron Meurer, Christopher P. Smith, Mateusz Paprocki et al. “SymPy : symbolic computing in Python”. In : *PeerJ Computer Science* 3 (2017), e103. doi : 10.7717/peerj-cs.103.

[Nov+08] John Novembre et al. “Genes Mirror Geography within Europe”. In : *Nature* 456 (2008), p. 98-101.

[ORa] John J. O'Connor et Edmund F. Robertson. *Andrei Andreyevich Markov*. MacTutor History of Mathematics Archive, University of St Andrews. https://mathshistory.st-andrews.ac.uk/Biographies/Markov/.

[ORb] John J. O'Connor et Edmund F. Robertson. *Benoit Mandelbrot*. MacTutor History of Mathematics Archive, University of St Andrews. https://mathshistory.st-andrews.ac.uk/Biographies/Mandelbrot/.

[ORc] John J. O'Connor et Edmund F. Robertson. *Euclid of Alexandria*. MacTutor History of Mathematics Archive, University of St Andrews. https://mathshistory.st-andrews.ac.uk/Biographies/Euclid/.

[ORd] John J. O'Connor et Edmund F. Robertson. *Felix Hausdorff*. MacTutor History of Mathematics Archive, University of St Andrews. https://mathshistory.st-andrews.ac.uk/Biographies/Hausdorff/.

[ORe] John J. O'Connor et Edmund F. Robertson. *Helge von Koch*. MacTutor History of Mathematics Archive, University of St Andrews. https://mathshistory.st-andrews.ac.uk/Biographies/Koch/.

[ORf] John J. O'Connor et Edmund F. Robertson. *Johann Carl Friedrich Gauss*. MacTutor History of Mathematics Archive, University of St Andrews. https://mathshistory.st-andrews.ac.uk/Biographies/Gauss/.

[ORg] John J. O'Connor et Edmund F. Robertson. *Karl Pearson*. MacTutor History of Mathematics Archive, University of St Andrews. https://mathshistory.st-andrews.ac.uk/Biographies/Pearson/.

[ORh] John J. O'Connor et Edmund F. Robertson. *Karl Theodor Wilhelm Weierstrass*. MacTutor History of Mathematics Archive, University of St Andrews. https://mathshistory.st-andrews.ac.uk/Biographies/Weierstrass/.

[ORi] John J. O'Connor et Edmund F. Robertson. *Stefan Banach*. MacTutor History of Mathematics Archive, University of St Andrews. https://mathshistory.st-andrews.ac.uk/Biographies/Banach/.

[ORj] John J. O'Connor et Edmund F. Robertson. *Wacław Sierpiński*. MacTutor History of Mathematics Archive, University of St Andrews. https://mathshistory.st-andrews.ac.uk/Biographies/Sierpinski/.

[Pag+99] Lawrence Page et al. *The PageRank Citation Ranking : Bringing Order to the Web*. Rapp. tech. 1999-66. Stanford InfoLab, 1999.

[Pea01] Karl Pearson. "On Lines and Planes of Closest Fit to Systems of Points in Space". In : *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 6e série* 2.11 (1901), p. 559-572.

[Pic90] Émile Picard. "Mémoire sur la théorie des équations aux dérivées partielles et la méthode des approximations successives". In : *Journal de Mathématiques Pures et Appliquées* 6 (1890), p. 145-210.

[Pik03] Rob Pike. *UTF-8 history*. Courriel du 30 avril 2003, reproduit sur le site de l'auteur. https://www.cl.cam.ac.uk/~mgk25/ucs/utf-8-history.txt. 2003.

[Poi08] Henri Poincaré. *Science et méthode*. Paris : Flammarion, 1908.

[RSA78] Ronald L. Rivest, Adi Shamir et Leonard Adleman. "A Method for Obtaining Digital Signatures and Public-Key Cryptosystems". In : *Communications of the ACM* 21.2 (1978), p. 120-126.

[Sha38] Claude E. Shannon. "A Symbolic Analysis of Relay and Switching Circuits". In : *Transactions of the American Institute of Electrical Engineers* 57.12 (1938), p. 713-723.

[Sha48] Claude E. Shannon. "A Mathematical Theory of Communication". In : *Bell System Technical Journal* 27 (1948), p. 379-423, 623-656.

[She18] Weixiao Shen. “Hausdorff dimension of the graphs of the classical Weierstrass functions”. In : *Mathematische Zeitschrift* 289 (2018), p. 223-266.

[Sie15] Wacław Sierpiński. “Sur une courbe dont tout point est un point de ramification”. In : *Comptes rendus hebdomadaires des séances de l'Académie des sciences* 160 (1915), p. 302-305.

[Sie16] Wacław Sierpiński. “Sur une courbe cantorienne qui contient une image biunivoque et continue de toute courbe donnée”. In : *Comptes rendus hebdomadaires des séances de l'Académie des sciences* 162 (1916), p. 629-632.

[Sin99] Simon Singh. *Histoire des codes secrets : de l'Égypte des pharaons à l'ordinateur quantique*. Traduit de l'anglais par Catherine Coqueret. Paris : JC Lattès, 1999.

[Spe04] Charles Spearman. “« General Intelligence », Objectively Determined and Measured”. In : *The American Journal of Psychology* 15.2 (1904), p. 201-292.

[The25] The Unicode Consortium. *The Unicode Standard, Version 17.0*. Unicode Consortium, South San Francisco. https://www.unicode.org/versions/latest/. 2025.

[TP91] Matthew Turk et Alex Pentland. “Eigenfaces for Recognition”. In : *Journal of Cognitive Neuroscience* 3.1 (1991), p. 71-86.

[Uni92] United States General Accounting Office. *Patriot Missile Defense : Software Problem Led to System Failure at Dhahran, Saudi Arabia*. Rapport GAO/IMTEC-92-26, Washington. 1992.

[Usi82] Zalman Usiskin. *Van Hiele Levels and Achievement in Secondary School Geometry*. Rapport du projet CDASSG. Chicago : University of Chicago, 1982.

[Yan11] H. Bahadir Yanik. “Prospective middle school mathematics teachers' preconceptions of geometric translations”. In : *Educational Studies in Mathematics* 78.2 (2011), p. 231-260.